DOCUMENT RESUME

ED 469 771 TM 034 534

AUTHOR	Zhang, Liru; Lau, Allen; Slinde, Jeffery
TITLE	Delaware Student Testing Program: Technical Report, 2001.
INSTITUTION	Delaware State Dept. of Education, Dover. Assessment and Accountability Branch.
PUB DATE	2002-07-00
NOTE	114p.
AVAILABLE FROM	For full text: http://www.doe.state.us.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC05 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; English; Language Arts; Mathematics; Psychometrics; *Reliability; Sciences; Social Studies; *State Programs; Test Construction; *Test Results; *Testing Programs; *Validity; Writing (Composition)
IDENTIFIERS	*Delaware Student Testing Program

ABSTRACT

The Delaware Student Testing Program (DSTP) is a mandated statewide assessment program that targets four core content areas. English language arts, including reading and writing, and mathematics are given to students in grades 3, 5, 8, and 10; science and social studies are given to students in grades 4, 6, 8, and 11. The DSTP reading and mathematics tests consist of two portions: items developed by Delaware educators that specifically measure the Content Standards and selected items from the Stanford Achievement Series, ninth edition. The DSTP science and social studies tests consist of Delaware-developed items only. This document reports technical characteristics of the 2001 DSTP in reading, writing, mathematics, science, and social studies. Validity evidence and reliability data of test scores are presented for each test to support the technical quality of the statewide assessment program. Empirical evidence is also available to provide additional technical information about the DSTP. The report contains these sections: (1) "Introduction"; (2) "Design and Validity of the DSTP"; (3) "Reporting DSTP Results"; (4) "Design and Application of Scaling and Equating"; (5) "Technical Characteristics of the DSTP"; and (6) "References." Eleven appendixes contain test specifications and additional details about test characteristics. (Contains 11 tables and 16 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 469 771

Delaware Student Testing Program

Technical Report - 2001

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

V. Woodruff

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034534

Assessment and Analysis Group
Assessment and Accountability Branch
Delaware Department of Education

July 2002

BEST COPY AVAILABLE

Officers of the Delaware Department of Education

Valerie A. Woodruff
Secretary of Education
Jennifer W. Davis
Deputy Secretary of Education

Robin R. Taylor, M. Ed.
Associate Secretary, Assessment and Accountability Branch
Mark A. Dufendach, Ed. D.
Associate Secretary, Finance and Administrative Services Branch
Nancy J. Wilson, Ph. D.
Associate Secretary, Curriculum and Instructional Improvement Branch

Wendy B. Roberts, Ph. D.
Director, Assessment and Analysis Group

Darlene J. Bolig, Ed. D.	Helen Dennis, M. Ed
Jeffery L. Fleming, M. S.	Katia Foret, Ph. D.
James F. Hertzog, M. Ed.	Nancy A. Maihoff, Ph. D.
Jon Manon, Ph. D., University of Delaware	
Joann F. Prewitt, M. A.	Julie A. Schmidt, Ph. D.
Carole D. White, M. B. A.	Liru Zhang, Ph. D.

Support Staff:

Elaner M. Brown	Krista D. Holloway
Barbara F. O'Neal	Erin L. Pieshala
Kimberly K. Rodriguez	Gail Truxon

State Board of Education
Joseph A. Pika, Ph.D., President
Jean W. Allen, Vice President
Robert J. Gilsdorf
Mary B. Graham, Esquire
Valarie Pepper
Dennis J. Savage
Claibourne D. Smith, Ph. D.

The Department of Education does not discriminate in employment or educational programs, services or activities, based on race, color, national origin, age, or handicap in accordance with state and federal laws. For more information about the DSTP, write to the Department of Education, Assessment & Accountability Branch, P. O. Box 1402, Dover, DE 19903-1402, or telephone (302) 739-4606.

This report is available on the WWW at http://www.doe.state.us under the Educational Resources link. Document Control No. 95-01/01/07/04

BEST COPY AVAILABLE

This report was prepared by Liru Zhang, Assessment and Analysis, in collaboration with Allen Lau, Harcourt Educational Measurement and Jeffery Slinde, Beck Evaluation & Testing Associate, Inc. The draft of the report was reviewed by the DSTP Technical Advisory Committee members, psychmetricians from the contractors, and related staff from the Department of Education.

Please contact Wendy Roberts, Director of Assessment and Analysis at (302) 739-6700 or by e-mail: wroberts@state.de.us if you have any questions regarding this report.

Table of Contents

Part One. Introduction

I. Delaware Student Testing Program (DSTP) 1
II. Organizations and Groups Involved 2
III. About This Report 3

Part Two. Design and Validity of the DSTP

I. Overview of Test Development 3

Ia. Delaware Content Standards 3
Ib. Developing Test Specifications 3
Ic. Item Development and Review 9
Id. Bias Review 10
Ie. Field Testing 11
If. Test Construction 12

II. Other Validity Evidence 12

IIa. Test Administration and Security 12
IIb. Inclusion Guidelines and Exemptions 13
IIc. Accommodations and Test Modification 13
IId. Student Questionnaire 14
IIe. Standard Setting for Science and Social Studies 16

Part Three. Reporting DSTP Results

I. DSTP Scores 18
II. DSTP Scoring 18
III. Reporting DSTP Results 19
IV. 2001 DSTP Scores 20

Part Four. Design and Application of Scaling and Equating

I. Design of DSTP Scale in Reading, Mathematics, Science, and Social Studies 20
II. Equating 25

IIa. Equating DSTP Reading and Mathematics 25
IIb. Equating DSTP Science and Soial Studies 26

III. 2001 Equating and Scaling Results 27

Part Five. Technical Characteristics of the DSTP

I. Reliability for the DSTP 28

Ia. Estimate of Reliabilities and Standard Error of Measurement 28
Ib. Correlations 29
Ic. Rater Consistency 29

II. Item and Test Statistics 30

Part Six. References 31

Appendix

Appendix A. Test Development Committees 32

Appendix B. Bias Review Committee 33

Appendix C. Technical Advisory Committee 34

Appendix D. Test Specifications for Reading 35

Appendix E. Test Specifications for Mathematics 36

Appendix F. Test Specifications for Science 37

Appendix G. Test Specifications for Social Studies 38

Appendix H. Frequency Distributions of Test Scores 39

Appendix I. Conversion Tables from Raw Scores to Scale Scores 40

Appendix J. Comparisons of Step-Values for Anchor Items by Test 41

Appendix K. Histogram Distributions of P-Values by Test and Year 42

List of Tables

Table 1. Statistics of Reading, Writing, and Mathematics Scores

Table 2. Statistics of Science and Social Studies Sub-Scores

Table 3. Percentage of Students in Each Performance Level by Grade and Test

Table 4. Reliability Coefficients of Test Scores by Grade and Test

Table 5. Correlation Matrix in Reading

Table 6. Correlation Matrix in Mathematics

Table 7. Correlation Matrix in Reading and Writing

Table 8. Correlation Matrix in Science

Table 9. Correlation Matrix in Social Studies

Table 10. Raters' Correlation and Raters' Agreement for Writing in 2001

Table 11. Summary of Item Statistics by Test and Grade

Part One. Introduction

I. Delaware Student Testing Program (DSTP)

The Delaware Student Testing Program (DSTP) is a mandated, statewide assessment program that targets four core content areas. English language arts, including reading and writing, and mathematics are given to students in grades 3, 5, 8, and 10; science and social studies are given to students in grades 4, 6, 8, and 11. The first two years' administrations, 1998 and 1999, serve as the baseline for determining the proficiency levels and student progress toward the standards in reading, writing, and mathematics; and the first two years' administrations, 2000 and 2001, serve as the baseline in science and social studies

The DSTP reading and mathematics tests consist of two portions: items developed by Delaware educators that specifically measure the Content Standards and selected items from the Stanford Achievement Series, 9th edition (SAT9) abbreviated version of reading comprehension; and mathematical problem solving for grades 3, 5, and 8 and mathematics for grade 10. The DSTP science and social studies consist of Delaware-developed items only.

Three types of item formats, multiple-choice, short answer, and extended constructed-response items, are used in reading and mathematics. Two scores, percentile rank (PR) on the SAT9 and standard-based scores (SBS) are reported at the state, school district, school, and individual levels. Percentile rank scores are based on the 30 multiple-choice items of the SAT9 sub-test; standards-based scores or scale scores are derived from the combination of selected SAT9 items of each sub-test and Delaware-developed items. The DSTP science and social studies tests include Delaware-developed items only in the formats of multiple-choice and short answer. Four sub-content areas are measured in science: inquiry, physical science, earth science, and life science and social studies: civics, economics, geography, and history, respectively. In addition to the standard-based scores (SBS) in science and social studies reported at the state, school district, school, and individual levels, the raw scores for each of the four sub-content areas are available for the use of Delaware educators.

Short answer and extended constructed-response items are scored by one trained rater from the contractor. To ensure the accuracy of scoring, about 10% of students' scores are checked by the team leader during the scoring process.

The DSTP writing test consists of a stand-alone writing prompt and a text-based writing task. For the standard-alone writing, students have approximately three hours for the completion of the essay, which includes directions, prewriting (20 minutes), first draft (30-45 minutes), and second draft (60 minutes), and a break time. The text-based writing is attached to a reading passage. Students must read the passage in order to respond to the text-based writing task. Students are encouraged to use prewriting for planning their writing. The stand-alone writing is scored by two trained raters and the

text-based writing is scored by one trained rater. The sum of the two scores for stand-alone writing is used as the stand-alone writing score on the 0-10 scale. Students' responses to the text-based writing are scored for writing using the same scoring rubric as for stand-alone writing on the 0-5 scale and for reading using the specific scoring rubric on the 0-4 scale. The total writing score is a composite score of the stand-alone and the text-based writing scores that is reported at the state, school district, school, and individual levels.

Student performance in reading, writing, mathematics, science, and social studies are reported in five performance levels, Distinguished, Exceeds the Standard, Meets the Standard, Below the Standard, and Well Below the Standard. The procedures for setting the cut-scores in science and social studies for grades 8 and 11 in the Spring and for grades 4 and 6 in the Fall of 2001 is briefly described in **Part Two: Design and Validity of the DSTP** – *IIe. Standard Setting for Science and Social Studies* of this report.

II. Organizations and Groups Involved

- *Harcourt Educational Measurement* has been the DSTP contractor since 1997. Their responsibilities include test development, training, test administration, data analysis, and reporting. The responsibilities of the sub-contractor, *National Computer System* (NCS), are data analysis and reporting.

- *Beck Evaluation & Testing Associate, Inc.*, a sub-contractor to *Harcourt Educational Measurement,* is responsible for standard setting and provides psychometric consultation.

- *The Test Development Committees* consist of Delaware teachers and content specialists (Appendix A). The committee members sign an annual contract with the Department of Education for item development, which may rotate in a 2 to 3-year cycle. All newly developed items are reviewed within the grade-cluster group or by another group of content specialists for content accuracy and standards alignment. The committee meets on a regular basis for training, reviewing/editing items, and discussing issues in item development. Based on the policy of the Department, 20% to 30% of the committee members are replaced each year.

- *The Bias Review Committee* consists of educators who have varied educational and cultural/ethnical backgrounds (Appendix B). The committee members carefully review each newly developed item in all content areas and reading passages for bias using the predetermined criteria and make recommendations for any modifications or re-writing the item.

- *The Technical Advisory Committee* consists of national experts in educational measurement and testing (Appendix C). The function of this committee is to advise the Department of Education to ensure that the DSTP provides a valid and reliable measure of student progress toward the Standards. The committee members review the DSTP design, the process of test development, and student data to address

concerns regarding the technical quality of the test and make recommendations for improvement.

- *The Benchmark Committees* are responsible for establishing anchor papers and assigning score point(s) to each anchor paper. The anchor papers are used for scoring short answer, extended response items, and writing prompt for the current year's DSTP. Each committee consists of 5 to 6 classroom teachers from the state and 1 to 2 members from the Test Development Committees.

III. About this Report

This document is prepared to report technical characteristics of the 2001 DSTP in reading, writing, mathematics, science, and social studies. Validity evidence and reliability data of test scores are presented in this report for each test to support the technical quality of the statewide assessment program. Empirical evidence is also available to provide additional technical information about the DSTP.

Part Two. Design and Validity of DSTP

I. Overview of Test Development

Ia. Delaware Content Standards

The standards-based educational reform began in Delaware in 1991. With the adoption of the rigorous content standards in English language arts, mathematics, science, and social studies in 1995, Delaware educators have continued the efforts to implement standards-based curriculum and assessment in order to meet the goals of improving achievement for all Delaware students. In 1997, the State Legislature passed the laws (Delaware Code, Title 14) that made the Delaware Student Testing Program the official measure of student progress toward the Delaware Content Standards. Student test scores are used as primary indicators of the accountability system.

Ib. Developing Test Specifications

The first step in test development is to develop the test specifications. Test content and skills measured by a test and distributions of emphasis are described in the test specifications with percentage or number of items under each category. Every test form of a given content area is developed based on the same test specifications across years so that student progress can be evaluated toward the standards over time. To ensure that the DSTP aligns with the Delaware Content Standards, the Test Development Committee worked, in conjunction with the content specialists from Harcourt Educational Measurement, to develop the test specifications for each test. Criteria used for determining the test content and skills, item types, and distribution of emphasis in the test specifications include the following:

- The importance of the content domain and skills specified in the Standards for a given grade;
- The performance level specified in the Standards for a given grade;
- The impact of the DSTP on curriculum and classroom instruction;
- The accessibility of the standards in a large-scale testing environment; and
- The item format that would measure the content and the performance level.

A. English language arts It is required in the English language arts Content Standards that all students in Delaware public schools should become effective readers, writers, listeners, viewers, and speakers. Due to the limits of large-scale testing, only reading and writing are assessed in the DSTP.

The reading assessment is designed to measure Standard 2 that students "construct, examine, and extend the meaning of literary, informative, and technical texts through listening, reading, and writing" and Standard 4 that students "use literary knowledge accessed through print and visual media to connect self to society and culture." Three types of reading passages, literary, informative, and technical, are used in the reading assessment. Three stances, determining meaning, interpreting meaning, and extending meaning, are used to measure the depth of reading comprehension.

- Questions in the determining meaning stance require the reader to demonstrate an overall understanding of the passage. The focus is on how the reader begins to make meaning of the text.
- Questions in the interpreting meaning stance require the reader to go beyond the initial understanding to develop an interpretation of the text. The reader goes beyond first impression to construct a more complete understanding of what has been read.
- Questions in the extending meaning stance require the reader to stand apart from the text and critically consider it. This stance involves critical examination, evaluation, and analysis.

The test specifications reflect the emphasis of types of reading passages and questions (stances) for each grade. For example, approximately 65% of the reading passages are literary, but only 15% are technical for grade 3; whereas for grade 10, 35% of the reading passages are literary and 40% are informative (Appendix D).

The writing assessment, consisting of stand-alone and text-based writings, is designed to measure Standard 1 "Use written and oral English for various purposes and audiences." For the stand-alone writing, students write an extended essay responding to a writing prompt; for the text-based writing, students write a short essay responding to a question about a passage in the reading assessment. Three discourses of writing tasks are used in

the DSTP writing assessment for the stand-alone writing: expressive, informative, and persuasive.

- Expressive (author-oriented)

  - Reveal self-discovery and reflection;
  - Demonstrate experimentation with techniques which could include dialogue;
  - Demonstrate experimentation with appropriate modes, which could include narration and description.

- Informative (subject-oriented)

  - Begin to address the needs of the audience;
  - Exhibit appropriate modes which could include description, narration, classification, simple process analysis, simple definition;
  - Conform to the appropriate formats, which could include letters, summaries, messages, and reports.

- Persuasive (audience-oriented)

  - Begin to consider the needs of the audience;
  - Communicate a clear-cut position on an issue;
  - Support the position with relevant information, which could include personal opinions and examples;
  - Exhibit evidence of reasoning.

B. Mathematics The mathematics assessment is designed to measure the Mathematics Standards. Multiple-choice (MC), short answer (SA), and extended constructed-response (ECR) items are used to measure the mathematics concepts and procedures in computation, estimation, and measurement; number sense; algebra; pattern and functions; geometry; and probability and statistics.

- Students will develop an understanding of Estimation, Measurement, and Computation by solving problems in which there is a need to measure to a required degree of accuracy by selecting appropriate tools and units; to develop computing strategies and select appropriate methods of calculation from among mental math, paper and pencil, calculators or computers; to use estimating skills to approximate an answer and to determine the reasonableness of results.

- Students will develop Number Sense by solving problems in which there is a need to represent and model real numbers verbally, physically and symbolically; to use operations with understanding; to explain the relationships between numbers; to apply the concept of a unit; and to determine the relative magnitude of real numbers.

- Students will develop an understanding of Algebra by solving problems in which there is a need to progress from the concrete to the abstract using physical model,

equations and graphs; to generalize number patterns; and to describe, represent and analyze relationships among variable quantities.

- Students will develop *Spatial Sense* and an understanding of Geometry by solving problems in which there is a need to recognize, transform, analyze properties of, and discover relationships between geometric figures.

- Students will develop an understanding of Statistics and Probability by solving problems in which there is a need to collect, appropriately represent, and interpret data; to make inferences or predictions; to present convincing arguments; and to model mathematical situations to determine the probability.

Mathematics concepts are measured on three cognitive processes: conceptual knowledge, procedural knowledge, and mathematical process (or problem solving) (Standards 1 to 4). The percentages of the cognitive processes are the same across grades, 40% for conceptual knowledge, 40% for procedural knowledge, and 20% for problem solving (Appendix E).

- Conceptual Knowledge involves the construction of fundamental mathematical ideas including the notions of a unit, counting, ordering, part vs. whole, mathematical operations, geometric figures, pattern, measurement, and chance. Conceptual knowledge deepens, as concepts are connected one to another and applied more widely.

- Procedural Knowledge involves the skills performance of a standardized routine. It tends to be most firmly held if based upon strong conceptual foundations. Not all-conceptual knowledge, however, is transformed into procedural knowledge. Only certain fundamental procedures need to be practiced to the point of fluency.

- While conceptual and procedural knowledge can be assessed independently, the application of these concepts and skills is described under the broad category of Mathematical Processes (also called Problem Solving). Both concepts and procedures may be called upon in non-routine situations that require problem solving. Connections between diverse mathematical concepts or between mathematics and another discipline can be assessed.

C. Science The science assessment is designed to measure the Science Standards. Eight standards cover core scientific concepts and critical skills under four sub-content areas (inquiry, physical science, earth science, and life science), which reflect the increasing complexity of science education that develop the capacity for life-long learning. The test specifications that are developed based on the Standards show varying emphases of each sub-content area from grade to grade (Appendix F). Multiple-choice and short answer items are used to measure students' knowledge and skills at different thinking levels.

- Science as Inquiry

The practice of science and the development of technology are critical pursuits of our society. These pursuits have involved diverse people throughout history and have led to continuous improvement in the quality of life and in our understanding of nature. Students will study the process of scientific inquiry and technology development and the history and context within which these have been carried out. In the science assessment, students will demonstrate their understanding skills to observe, experiment, and analyze data in scientific sittings.

- Physical Science

  - Materials and Their Properties: Students will develop a basic understanding of the structure and properties of materials. They will also experience and learn the process by which materials are changed and how the uses of materials are related to their properties.

  - Energy and Its Effects: Students will study, discuss, and learn the factors that govern the flow of energy throughout the universe, the transformation of natural resources into useful energy forms, and the conservation of energy during interaction with materials.

- Earth Science

  - Earth in Space: Students will learn that even though the distributions and types of materials differ from planet to planet, the chemical composition of materials is identical and the same laws of science apply across the universe.

  - Earth Dynamic Systems: Students will study and learn to identify components of the various Earth systems and understand the changes and patterns that result from interactions within and between these systems.

- Life Science

  - Life Processes: Students will learn how living organisms use matter and energy to build their structures and conduct their life processes. They will learn the mechanisms and behaviors used by living organisms to regulate their internal environments and to respond to changes in their surroundings. Students will also study how knowledge about life processes can be applied to improving human health and well being.

  - Diversity and Continuity of Living Things: Students will study how living things reproduce, develop, and transmit traits, and how theories of evolution explain the unity and diversity of species found on earth. Student will also study how knowledge of genetics, reproduction, and development is being applied to improve agriculture and human health.

- Ecology: Students will acquire a basic understanding of the structure of ecosystems and how they function and change. They will also study how humans can apply scientific and technological knowledge about ecosystems in making informed decisions about the use of natural resources.

D. Social Studies There are 16 concepts covered equally with 4 concepts for each of the four sub-content areas: civics, economics, geography, and history for each grade-cluster in the Social Studies Standards. The complexity of performance level, however, increases at each succeeding grade-cluster. For example, Civics Standard One indicates "Students will examine the structure and purpose of government with specific emphasis on constitutional democracy." It is required that "students will understand that governments have the power to make and enforce laws and regulations, levy taxes, conduct foreign policy, and make war" for the grade-cluster 4-5; while "students will analyze the ways in which the structure and purposes of different governments around the world reflect differing ideologies, culture, values, and histories" for the grade-cluster 9-12. The assessment is designed to measure the Content Standards for social studies. The Test Development Committee recognized that all the four sub-content areas are equally important to all grade-clusters in social studies and believed that the integration of the four sub-content areas in social studies is important in classroom instruction. Thus, all the standards are specified with equal emphasis in the test specifications across grades, except one in history that is very difficult to be measured in large-scale testing (Appendix G).

- Civics

Students will examine the structure and purposes of governments with specific emphasis on constitutional democracy; understand the principles and ideas underlying the American political system; understand the responsibilities, rights, and privileges of United States citizens; and develop and employ the civic skills necessary for effective, participatory citizenship.

- Economics

Students will analyze the potential cost and benefits of personal economic choices in a market economy; examine the interaction of individuals, families, communities, businesses, and governments in a market economy; understand different types of economic systems and how they change; and examine the patterns and results of international trade.

- Geography

Students will develop a personal geographic framework, or "mental map", and understand the uses of maps and other geo-graphics; develop a knowledge of the ways humans modify and respond to the natural environment; develop an understanding of the diversity of human culture and the unique nature of places; and develop an understanding of the characters and use of regions and the connections between and among them.

- History

Students will employ chronological concepts in analyzing historical phenomena; gather, examine, interpret, and analyze historical data; and develop historical knowledge of major events and phenomena in world, United States, and Delaware history.

Ic. Item Development and Review

Item Development Newly-developed items and scoring rubrics for short-answer and extended constructed –response questions are reviewed and edited by the grade-cluster group of the Test Development Committee or by the Advisory Committee for content accuracy and standards alignment. Every new item written by item writer is also reviewed and edited by the chair of the Test Development Committee and the content specialists from Harcourt Educational Measurement. Those formal editorial reviews provide the item writers with inputs for item improvement. Accepted edited items, then, are ready for bias review and then for the field test. The Department provides a week-workshop for all item writers in summer for training and item development. Small-scale try-out of newly developed items, whenever it applied, is an important way to get feedback from students' responses. The common criteria used for item review are listed below with specific criteria used for content review of items.

- Content accuracy
- Alignment to content standards
- Appropriate content to grade level
- Appropriate scoring rubrics for short answer and extended-constructed items
- Correct answer for multiple-choice items
- Appropriate item format to item content
- Clarity or no ambiguity
- Appropriate reading level to grade level

The specific criteria are used by the Test Development Committees in Reading and Writing for content review:

- Interesting topic of reading passages and writing prompt
- Attachment of items to the reading passage
- Use of EDL Core Vocabulary to check the readability
- Accuracy of wording

The specific criteria are used by the Test Development Committee in Mathematics for content review:

- Accuracy of formula, figures, and graphics
- Calculator-dependent or calculator-independent items are in the right session

The specific criteria are used by Test Development Committee in Science for content review:

- Developmental appropriateness of items
- Important topics rather than recalling detailed facts
- Accuracy of graphics

The specific criteria are used by the Test Development Committee in Social Studies for content review:

- Accuracy of graphics
- Alignment of graphics to the items

B. Review SAT9 Items To determine the alignment of the SAT9 reading and mathematics tests to the Delaware Content Standards, in late 1997, a Review Committee consisting of Delaware content specialists reviewed the SAT9 reading comprehension and SAT9 problem solving in mathematics item by item, coded each item with the Standards, and then classified items into two categories, items that measure the Standards and items that do not measure the Standards. The results of the alignment show that there are 27, 24, 27, and 24 items for grades 3, 5, 8, and 10, respectively, in SAT9 reading comprehension; 29 items for grades 3 and 5 and 30 items for grades 8 and 10 in SAT9 mathematics problem solving measure the corresponding Standards. To receive a reliable national comparison, all 30 SAT9 items in reading and mathematics are administered to students in grades 3, 5, 8, and 10 under standardized, timed testing conditions of approximately 30 minutes each. But only selected SAT9 items combined with Delaware-developed items to derive a standards-based score determine student progress toward the Delaware Content Standards.

Id. Bias Review

Regardless of the purpose of testing, fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure (Standards for Educational and Psychological Testing, 1999, p.74). Judgmental methods for review of test items are often supplemented by statistical procedures for identifying items on tests that function differently across identifiable sub-groups of examinees in large-scale test settings (Standards for Educational and Psychological Testing, 1999). In Delaware, sensitivity to item bias has been built into the process of test development since 1997. Items that pass the content review are submitted to the Bias Review Committee for bias review before the field test. Every newly developed item is reviewed using the predetermined criteria. The chairs of the Test Development Committees may attend the bias review meeting to answer questions only. The bias Review Committee members identify items that may contain stereotypes (e.g., sexism, racism), irrelevant constructs that pose particular difficulty for one sub-group (e.g., Limited English Proficient students), and/or biased content against different subgroups (e.g., gender, racial/ethnic, religion, socioeconomic status, geographic location). The Committee makes suggestions for modifying or re-writing flagged items.

To ensure that all students be given a comparable opportunity to demonstrate their standing on the construct(s) the DSTP is intended to measure, the Department of Education and Harcourt Educational Measurement conducted a pilot Differential Item Functioning (DIF) analysis in 2001. Two procedures, Mantel-Haenszel (MH) and Simultaneous Item Bias (SIBTEST) procedures were employed using the 1999 DSTP reading data. *Mantel-Haenszel (MH) Procedure* is the most commonly used non-IRT methodology for detecting item bias developed by Mantel and Hanenszel in 1959. The 2x2 contingency table is used for a specific level to compare how frequently examinees from the reference group (i.e., Whites, males) and those from the focal group (i.e., Blacks, females) pass or fail a particular item. This comparison is repeated for the remaining score levels, and the results are averaged across the various performance levels. MH employs two criteria for determining bias: (a) the statistical significance of χ^2; and (b) the magnitude and direction of Δ (Nandakumar, Glutting, & Oakland, 1993). Positive values of Delta indicate that the item is relatively easier for the focal group. Negative values indicate that the item is easier for the reference group. *SIBTEST* is one of the recently developed IRT-based methodologies for detecting DIF by Shealy and Stout (1993a, 1993b). SIBTEST can be used either to detect item bias/DIF or to detect test bias/DTF simultaneously. The null hypothesis of no DIF is rejected with error rate α if the value of *B* exceeds the upper $100(1 - \alpha)$th percentile point of the standard normal distribution. βu is the statistic used to estimate the amount of unidirectional DIF. For example, a βu value of 0.1 indicates that the average difference between the expected total test scores for reference and focal group examinees of similar ability is 0.1. Positive values of βu indicate DIF against the focal group (i.e., females, Blacks) and negative values βu of indicate DIF against the reference group (i.e., males Whites) (Nandakumar, 1993). The following findings are based on the initial review of the results:

- The DIF results are generally consistent between the two procedures for grade 8.
- The results are more consistent between the two procedures for detecting DIF items for gender groups than for racial groups.
- It seems that SIBTEST is more sensitive than MH in detecting DIF items, especially for short answer and extended constructed-response items.

In 2001, the Mantel-Haenszel (MH) procedure was applied to all field test items as supplement to the judgmental procedure for examining item bias and selecting items. It is intended that the DIF analysis will be applied to all core items starting in 2002.

Ie. Field Testing

Test items that pass the content review and bias review are formatted into field tests for the purpose of generating statistical characteristics of the items. Four to six field test forms per grade per test are embedded into the operational test form for each DSTP administration. The test forms are spiraled within classroom and school. The field test design is due to the following reasons:

- Minimize sampling errors;

- Minimize errors in item statistics because of student motivation;
- Minimize interruption of regular classroom instruction;
- Reduce the budget for testing; and
- Easy test administration.

If. Test Construction

After each year's test administration, the Test Development Committees make recommendations on the replacement of about 30% of the test items from the operational test forms. Professional judgment, in addition to statistical evidence, provides the basis for item selection. Objective data from the field tests are used primarily for evaluating the technical characteristics of test items, such as item difficulty, item discrimination, and the strengths and weaknesses of distractors for multiple-choice items and scoring rubrics and samples of student responses for short answer and extended constructed response questions. The newly selected items, in conjunction with the remaining items from the previous year's test form, must match the predetermined test specifications precisely. Then, the newly assembled test forms are reviewed and approved by the Test Development Committee. The process of item selection and test assembly using the item statistics from the field test and the test specifications ensure that each newly developed test form demonstrates the desired psychometric characteristics and provides supportive construct validity evidence.

II. Other Validity Evidence

Additional validity evidence for the 2001 DSTP is summarized in the following section, such as DSTP administration, accommodations provided for special student populations, a summary of the 2001 Student Survey Questionnaire, and standard setting in science and social studies for grades 4, 6, 8, and 11.

IIa. Test Administration and Security

The *Test Coordinator's Handbook* provides the guidelines for planning and managing the DSTP administration for District and School Test Coordinators. The *Directions for Administering* by grade and test provided specific directions for test administrators from room arrangement, scheduling and timing for subtests, and preparing students to testing students of the special populations. A comprehensive training session jointly conducted by the Delaware Department of Education and Harcourt Educational Measurement was scheduled in the fall and the spring before the testing week for School and District Test Coordinators.

The SAT9 reading comprehension and mathematics problem solving tests were administered under standardized, timed testing conditions. Three untimed sessions in reading and two 60minute sessions in mathematics were given for the Delaware portion in separate days. Commonly used mathematical formulas for grades 8 and 10 were provided as a reference during testing. Calculators (graphing calculators only for grade 10) were allowed for one session of the Delaware portion. Two 65-minute sessions were

BEST COPY AVAILABLE

given for science and social studies. Students might take longer time to complete the tests.

In writing, both stand-alone and text-based writings were untimed. The stand-alone writing took approximately 3 hours, including pre-writing, first draft, and the second draft. Only the second draft was scored. A checklist was provided and dictionaries were available for all students. Started in 2001, students were provided with an additional blank page used for pre-writing for the text-based writing and the format of the text-based writing was changed to look more like the format of the stand-alone writing.

The Test Security Guidelines indicate that photocopying of all or any part of a test booklet was ***strictly prohibited*** and all known violations of the Delaware Department of Education's regulations for test security should be reported immediately. As usual all test booklets were secure materials. Each test booklet was individually numbered with a unique bar code label. The District/School Test Coordinators were required to document the receipt of secured materials, check the lists of students, and return all test materials to Harcourt Educational measurement for scoring by schedule.

IIb. Inclusion Guidelines and Exemptions

The DSTP is a statewide, mandated assessment that is intended to include all of the public school students in Delaware. However, students in the life-skills curriculum are exempted from the DSTP under the Individuals with Disabilities Education Act (IDEA) or Section 504 of the Rehabilitation ACT. These students were assessed using the Delaware Alternate Portfolio Assessment (DAPA). Limited English proficient (LEP) students who have been in Delaware schools less than one year may be exempted from the DSTP one-time only. The decision to exempt was made on the individual basis from professional judgments of the LEP teacher, the principal, and/or the LEP contact person. Corresponding documentation for the exemption was required.

IIc. Accommodations and Test Modification

A variety of accommodations and test modification strategies have been implemented for students with disabilities and limited English proficient students (LEP). It is important to recognize that accommodated testing conditions and test modifications do not change the construct nor affect the psychometric characteristics of the assessment.

In 2000, two Task Forces, the DSTP Disability Task Force consisting of special education teachers, administrators, school psychologists, speech therapists and the Language Minority Task Force consisting of ESL, LEP, and bilingual teachers, reviewed and discussed related federal policies, Delaware regulations, and the existing accommodations for special education and LEP students. To include as many students as possible in the statewide assessment and meet the needs of students from special groups, the two Task Forces recommended the policies for exemption from the DSTP, eligibility for alternative assessment for special education students, and corresponding accommodations for special education students and LEP students. The Technical

Advisory Committee reviewed the accommodations and communicated with the representatives of the Task Force. The discussion primarily focused on whether the modified testing conditions changed the test construct and thus, affected the comparability of test scores. Recommendations for changes were made accordingly for aggregated and non-aggregated accommodations.

A Braille form was available for blind students except the items depending on complicated graphics (Items not applicable for blind students are omitted). Spanish versions of the DSTP were translated directly from the English versions in mathematics, science, and social studies for limited English proficient, Spanish-speaking students. Test scores on both modified forms are aggregated. Adjusted scale scores were developed for test scores on the Braille form because some items were omitted from the original test form.

IId. Student Questionnaire

In 2001 students in grades 4, 6, 8, and 11 were given a survey questionnaire along with the DSTP science and social studies tests. The survey questions were classified into three categories: Opportunity to Learn, Science, and Social Studies (For administration reason, students in grade 8 missed the five questions in Opportunity to Learn). Students' responses to the survey questions were analyzed by grade, gender, racial/ethnic group, and the performance level of a given test. The survey results provide additional validity evidence to support the design of the statewide assessment that measures student progress toward the Delaware Content Standards, the impacts of standards-based assessment on teaching and learning, and the connections between curriculum and student performance on the DSTP. The primary findings from the survey are presented in the following section. For details of the results, please see the 2001 Administration - State Summary of Student Questionnaire Report.

- About 55% and 47% of the students in grades 4 and 6, respectively, talked about what they had learned in school with someone at home almost every day, but only 20% of the students in grade 11 discussed their studies at home as frequently as their peers in lower grades. Across grades, a similar pattern is observed that more high-scoring students than low-scoring students in science frequently talked about what they had learned in school with someone at home, for example in grade 6, 59% of the students in Level 5, 50% Level 4, 47% Level 3, 44% Level 2, and 48% Level 1.

- According to the survey, 26% of the 4th graders, 36% of the 6th graders, and 32% of the 11th graders spent one hour or more on their homework and 60%, 47%, and 30% of the students in grades 4, 6, and 11, respectively, spent about a half hour on their homework per school day. Across the grades, more high-achieving than low-achieving students spent more time on their homework; while more low-achieving than high-achieving students reported no homework or had rarely done their homework. For example in grade 11, 49% of the students in Level 5, 45% Level 4, 38% Level 3, 30% Level 2, and 21% Level 1 spent more time on their homework; while 14%, 11%, 7%, 6%, and 3% from Level 1 to Level 5, respectively, did not have or did not do their homework.

- On the average, 39% of the students in grade 4, 45% in grade 6, and 30% in grade 11 spent two hours or more in watching television each school day. The data also show that more low-achieving students than high-achieving students spent more time watching television. For example in grade 6, 45% of the students in Level 1, 51% Level 2, 45% Level 3, 35% Level 4, and 26% Level 5 watched television for over two hours per school day.

- Less than one-third of the students in grades 4 (28%), 6 (18%), 8 (26%), and 11 (18%) actually used scientific equipment in their classes almost every day, 25%, 40%, 42%, and 26% in the four grades, respectively used scientific equipment once or twice a week in their science classes. The data indicate that more high-scoring than low-scoring students used scientific equipment more frequently; whereas more low-scoring than high-scoring students rarely used the equipment in their science classes. For example in grade 8, 19% of the students in Level 1, 13% Level 2, 8% Level 3, 5% Level 4, and 6% Level 5 reported that they have never or hardly ever used scientific equipment in their classes.

- Seventy-seven percent of the students in grade 4, 79% in grade 6, 74% in grade 8, and 56% in grade 11 believed that the concepts and knowledge they had learned in science classes helped them or somewhat helped them understand the world better.

- The survey results show that 74% of the students in grade 4, 71% in grade 6, 60% in grade 8, and 42% in grade 11 felt that their science classes had prepared them to do well on the DSTP science test. Moreover, 80% of the students in grade 4, 73% in grade 6, 63% in grade 8, but only 22% of the students in grade 11 reported that they had tried very hard to do well on the science test.

- Forty-two percent of the students in grade 4, 46% in grade 6, 41% in grade 8, and 30% in grade 11 reported that their teachers had asked them to apply the concepts and knowledge they learned in their social studies classes to solve real life problems every time or most of the time. About one third of the students across the four grades reported that their teachers had asked them to use primary sources, such as documents, diaries, and artifacts, at least most of the time in their social studies classes.

- Over 40% of the students in grades 6 (43%), 8 (46%), and 11 (40%) reported that their teachers had asked them in most of their social studies classes to explain why there are often different interpretations of the same event. Based on the survey, more high-scoring students than low-scoring students explained the reasons for different interpretations of the same event every time, for example in grades 8 (24%, 19%, 16%, 13%, and 12% from Level 5 to Level 1, respectively) and grade 11 (28%, 20%, 19%, 13%, and 10% from Level 5 to Level 1, respectively).

- The survey data show that the instructional time that social studies teachers spent in each or most their classes teaching the four sub-content areas of social studies varied. In grade 4, teachers spent 38% of the instructional time for geography, 25% for civics

& government, 25% for economics, and 41% for history; in grade 6, teachers spent 40% of the instructional time for geography, 27% for civics & government, 26% for economics, and 59% for history; in grade 8, teacher spent 29% of the instructional time for geography, 56% for civics & government, 42% for economics, and 73% for history; in grade 1, teachers spent 21% of the instructional time for geography, 40% for civics & government, 33% for economics, and 60% for history.

- The survey results provide evidence to support the connections between the instructional time teachers spent for teaching the four sub-content areas and student performance on the DSTP social studies test. For example, in grade 6, more high-scoring students than low-scoring students reported that their teachers had taught geography in most of their social studies classes (62%, 52%, 41%, 36%, and 38% from Level 5 to Level 1, respectively); whereas more low-scoring students than high-scoring students reported that their teachers had rarely or hardly ever taught geography (2%, 7%, 13%, 20%, and 20% from Level 1 to Level 5, respectively). Similarly in grade 8, more high-scoring than low-scoring students reported that their teachers had taught civics & government in most of the classes (69%, 66%, 62%, 52%, and 43%, respectively); whereas more low-scoring than high-scoring students reported that their teachers had rarely or hardly ever taught this content area (3%, 5%, 6%, 9%, and 15%, respectively).

IIe. Standard Setting for Science and Social Studies

Using the two-year data in science and social studies (2000 and 2001) as the baseline, the standard setting was conducted in the spring for grades 8 and 11 and in the fall for grades 4 and 6 in 2001. To set fair and meaningful performance standards, a representative committee was organized for each grade of a given test. The Standard Setting Committees, consisting of classroom teachers, educators, administrators, parents, and representatives from governor's office, legislature, educational organizations and business community from throughout the state, represented diverse background and multi-cultural/ethnical groups.

The Item Mapping procedure (also called the Bookmark procedure) was applied to set the cut-scores in science and social studies. This approach requires a group of judges to examine a book of test items arranged from the easiest to the most difficult one and insert the "bookmarks" at the items they believe most strongly define where a cut-score should be placed. Each standard setting session took approximately one half-day for training on the instrument and one half-day for each of the three rounds of judgments. Impact data was also provided to help committee members make decisions. The primary responsibility of the Standard Setting Committees was to set two cut-scores, one was used to identify students who meet the standard and one was used to identify students whose performance exceed the standard.

To evaluate the procedure for standard setting, a survey was given to the Standard Setting Committee members. The results of the survey show that:

- About 90% (91% for grades 4 and 6; 87% for grades 8 and 11) of the committee members believed that the training was adequate or somewhat adequate in preparing them to make judgments about the performance levels.

- 81% of the members from the grades 4 and 6 committees felt high and relatively high confidence about the cut-scores for Exceeds/Meets the Standard; but only 34% of the members from the grades 4 and 6 committees felt the same way.

- 81% of the members from the grades 4 and 6 committees felt high and relatively high confidence about the cut-scores for Meets/Below the Standard; while 57% the members from the grades 8 and 11 committees felt the same way.

- Over 90% (90% for grades 4 and 6; 91% for grades 8 and 11) of the committee members reported that they had adequate opportunities to address their professional opinions during the process; and

- About 70% (75% for grades 4 and 6; 78% for grades 8 and 11) of the committee members believed that the performance levels were set based on professional judgments rather than outside influences.

Following the standard setting sessions, the Department of Education reviewed the two cut-scores per grade recommended by the committee and made minor adjustments if necessary. The standard error was utilized as the maximum threshold for the adjustment in order to provide consistency of cut-scores across grades and tests. The two years' impact data was also carefully reviewed for adjustments within a content area as opposed to set equal distances across grades on the same scale. As the two primary cut-scores were confirmed, the Department of Education recommended two additional cut-scores using the standard error for calculation, one was used to differentiate Below and Well Below the Standard and one was used to differentiate Exceeds the Standard and Distinguished level. The cut-scores were approved by the State Board of Education in February for grades 8 and 11 and in September for grades 4 and 6. The tables below show the cut-points from raw scores to scale scores by grade and test in 2001:

Cut-Scores for 2001 Science

Level	GR 4		GR 6		GR 8		GR 11	
	Raw	Scale	Raw	Scale	Raw	Scale	Raw	Scale
Well Below								
Below	23	286	20	285	20	280	18	282
Meets	33	300	30	300	30	300	28	300
Exceeds	51	325	47	325	43	325	44	325
Distinguished	57	336	53	335	49	338	50	335

Cut-Scores for 2001 Social Studies

Level	GR 4		GR 6		GR 8		GR 11	
	Raw	Scale	Raw	Scale	Raw	Scale	Raw	Scale
Well Below								
Below	25	285	21	286	19	282	17	276
Meets	35	300	31	300	29	300	28	300
Exceeds	51	325	49	325	45	325	41	325
Distinguished	57	337	55	335	51	335	47	337

Part Three. Reporting DSTP Results

I. DSTP Scores

In addition to the performance levels, two scores were reported for reading and mathematics and one score reported for writing, science, and social studies. Instructional comments were provided in reading, mathematics, and writing for the use of classroom instruction.

- Percentile rank (PR) based on the Stanford Achievement Series, 9th edition (SAT9) abbreviated version reported in reading comprehension and mathematical problem solving for grades 3, 5, and 8 and mathematics for grade 10;

- Standards-based score (SBS), a composite scale score of selected SAT9 items and Delaware-developed items, reported in reading and mathematics;

- A total writing raw score was reported. In addition, the scores on text-based writing and stand-alone writing were available for educators;

- Standard-based score (SBS), a scale scores based on Delaware-developed items only, reported in science and social studies. The raw scores of inquiry, physical science, earth science, and life science in science; and the raw scores of civics, geography, economics, and history in social studies were available for educators; and

- Student performance in reading, writing, mathematics, science, and social studies were reported in five levels, Distinguished, Exceed the Standard, Meets the Standard, Below the Standard, and Well Below the Standard, using predetermined cut-scores.

II. DSTP Scoring

Multiple-choice items were scored electronically; short answer, extended constructed-response items and students' essays were scored by trained readers at the Performance Assessment Scoring Center (PASC) of Harcourt Educational Measurement using pre-developed scoring rubrics. Students' essays on the stand-alone writing prompt were scored by two readers using the holistic rubric on a 5-point scale; students' responses to

the short-answer and extended constructed-response items and essays on the text-based writing task were scored by one reader only. About 10% of students' scores for each test at each grade level were examined by the team leader of scoring to determine the accuracy of scoring. Rater's consistency (or called rater's reliability) for writing assessment is discussed in **Part Five. Technical Characteristics of the DSTP** of this report.

The Performance Assessment Scoring Center (PASC) at Harcourt Educational Measurement was established in 1988. The preliminary criteria for recruiting and screening raters require a four-year college education and a writing sample, followed by an intensive introductory training workshop. In order to join the general pool of readers, the candidates completed a one-day general workshop for each subject area. All readers and team leaders who scored DSTP writing assessment and constructed-response items in reading, mathematics, science, and social studies participated in a project specific training before working on the actual project. In the process of scoring, the Scoring Director provided a training to ensure that readers became expert with the specific test at the grade level and worked closely with the team leaders and raters to monitor the accuracy and consistency of scoring.

Before anchor pulling took place, PASC Scoring Directors and Team Leaders studied the writing prompts, constructed-response items, and scoring rubrics thoroughly. They reviewed students' responses and pulled out range papers that represented the full range of quality as described in the rubric. Range papers were then sorted from low, medium, to high. They also identified problem papers, such as off topic and invalid responses.

The Benchmark Committee consisted of 5-6 Delaware teachers and 1-2 Test Development Committee members, one for each test at each grade level for the 2001 DSTP administration. The responsibilities of this committee was to (1) establish anchor papers that would be used to score students' responses; (2) assign score point(s) to each anchor paper; and (3) establish training sets for scoring.

Each selected paper was scored by all the committee members individually followed by group discussion. The iterative process of reading, charting, and discussing was designed to achieve three goals: (1) to establish virtual agreement on each paper; (2) to identify papers that were on the line between two adjacent scores and force the clarification of that line; and (3) to allow committee members to justify their scores. Complete agreement on the score assigned to each anchor paper was expected. The content experts from Harcourt Educational Measurement, then, reviewed the anchor papers and practice sets across all items in a content domain and across grades in a subject area to ensure consistent decisions and consistent application of scoring rubrics for adjustment. Training materials for scoring were prepared as well from the results of anchor paper pulling during the process. Approval was made by the Department of Education to finalize the anchor papers and training materials for scoring.

III. Reporting DSTP Results

DSTP results were reported at the individual, school, school district, and state levels according to predetermined aggregation rules in 2001. Students who were tested with non-aggregated accommodations, from Intensive Learning Centers (ILCs), did not meet the attempt requirements or did not receive a valid score of a given test were excluded from the summary at the school, district, and state levels. With the increasing use of technology, the DSTP Online Report provided the great opportunities for educators, classroom teachers, and general public to review the test results at various levels, disaggregated data, compare students' performance across years, and generate their own reports to improve teaching and learning. Teachers may also track their students' previous records from different schools. A State Summary Report was prepared as a DSTP document, including aggregated results, disaggregated results, and the performance for ILC schools. A State Writing Assessment Report was published to provide additional information in writing assessment, such as cross-year comparisons by grade, frequency distributions of the stand-alone and text-based writing by grade.

IV. 2001 DSTP Test Scores

Descriptive statistics of the SAT9 NCE scores and standard-based scores (SBS) in reading and mathematics, raw scores in writing, standard-based scores in science and social studies, and the percentage of student in each performance level in all five content areas are summarized in Table 1 to Table 3. The frequency distributions of scale scores in reading, mathematics, science, and social studies and raw scores in writing are presented in Attachment H. Both descriptive statistics and frequency distributions are based on all students who took the 2001 DSTP and received a valid score of a given test.

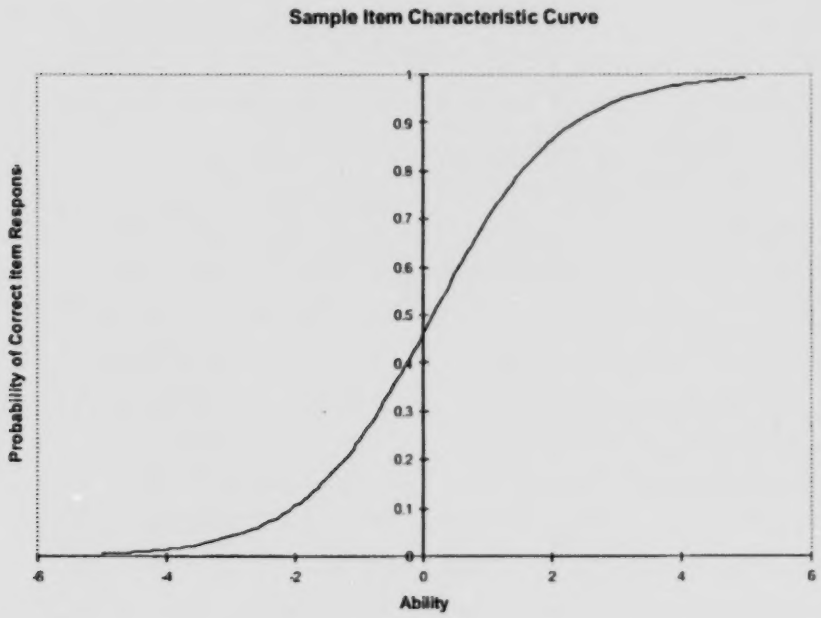
Part Four. Design and Application of Scaling and Equating

I. Design of DSTP Scale in Reading, Mathematics, Science, and Social Studies

The Rasch measurement model was used to develop the scale for each of the Delaware Reading, Mathematics, Science, and Social Studies assessments. This model has proven to be robust and sufficient for meeting the measurement needs of many large-scale, high stakes assessment programs. In general, the Rasch model assumes that the probability that a student will answer an item correctly is a function of the latent trait that underlies performance on the assessment and the difficulty of the item. The underlying trait, usually referred to as ability, is nothing more than what the assessment is designed to measure. The Rasch model is a mathematical function that relates the item score, or raw score, to the student achievement level or ability. Only item difficulty and person ability are used to define this mathematical function. It is the only Item Response Theory (IRT) model in which the student's raw score, the number of items answered correctly on the test, is a sufficient statistic—the only piece of information relevant for judging the rank ordering of a student on the ability continuum.

The most basic expression of the Rasch model is in the Item Characteristic Curve (ICC). A sample ICC is given in Figure 1. An Item Characteristic Curve is a mathematical function that relates the probability of a correct response to an item across the ability continum. The probability of getting a correct response is bounded by 1 (certainty of a correct response) and an incorrect response is bounded by 0 (certainty of an incorrect response). The ability scale is, in theory, unbounded and can range from $-\infty$ to $+\infty$. In practice, the ability scale ranges from approximately –4.00 to +4.00 logits for heterogeneous ability groups. A logit (natural log odds of a correct response) of zero typically represents "average" ability.

Figure 1



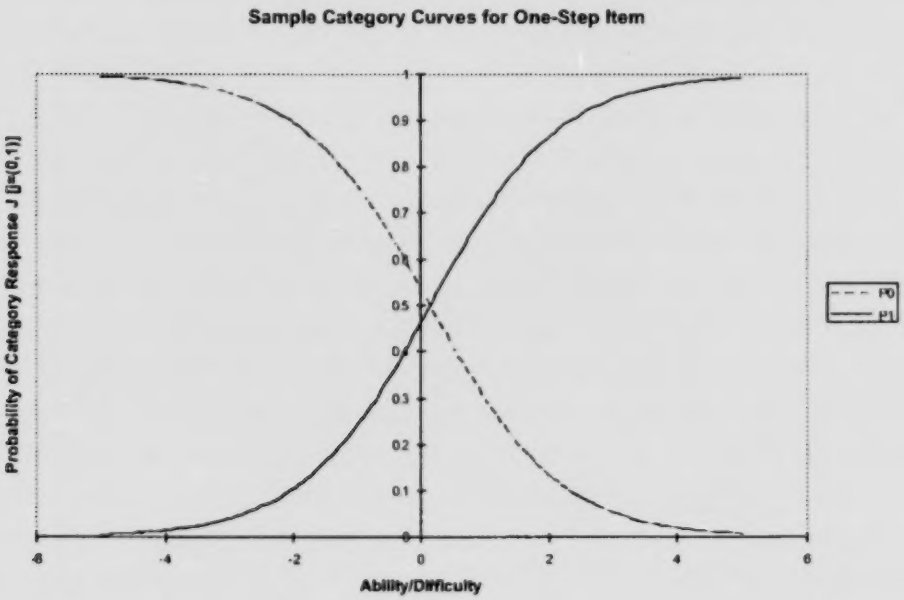
In Figure 1, a person whose ability falls at -1 on the ability (horizontal) scale has a probability of about 24% of answering the item correctly. Another way of expressing this is that if we have a group of 100 students, all of who have an ability of -1, we would expect about 24% of them to answer this item correctly. Similarly, a person whose ability was at +1 would have about a 70% chance of getting the item right. Thus, a person whose ability is above average is more likely to answer the item correctly than is one whose ability is below average. This makes intuitive sense and is the basic formulation of Rasch measurement for test items having only 2 possible categories (i.e., right or wrong).

To extend the formulation, consider that the Item Characteristic Curve shown above represents the Rasch expression that relates a person's ability to the *probability of a*

correct response to a given item. One might ask what sort of curve would represent the other possible condition, that of answering the item incorrectly. Intuitively, it would seem that if one has a probability of 70% of getting the answer right at an ability level of 1, then the probability of getting it wrong is 30%; at -1 on the ability scale, the probability of answering incorrectly is 76%. Thus, the less ability one has, the more likely he or she is to answer a test item incorrectly. This relationship, the *probability of an incorrect response,* is depicted by adding the second curve in Figure 2.

The point at which the two curves cross represents the ability level at which a person is just as likely to answer the item incorrectly as he or she is to answer it correctly. In other words, the probability of a correct (or incorrect) answer is 50%. This corresponds to the Rasch (logit) difficulty of a dichotomously scored item (e.g., multiple-choice item). The Rasch difficulty of a dichotomous item can also be referred to as the step value in going from a score of 0 to a score of 1.

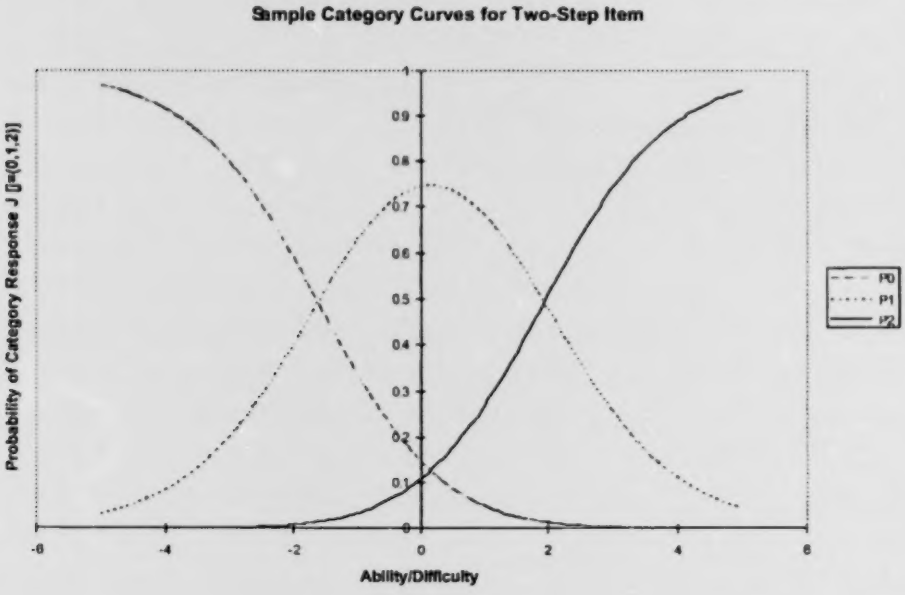
Figure 2



The description of the Rasch model so far has focused on multiple-choice items. But the DSTP reading, mathematics, science, and social studies assessments contain both multiple choice and constructed-response questions (e.g., short answer and extended constructed-response questions). With constructed-response items, students write their own response to the question. The student responses are scored in more than two (right/wrong) categories. The Rasch dichotomous model merges with the Partial Credit Model (PCM) by using additional response categories. Suppose that rather than scoring items as completely wrong or completely right, we add a category representing answers that, though not totally correct, are still clearly not totally incorrect. That is, the item is

scored in one of three categories; category 1 with score = 0, category 2 with score = 1, and category 3 with score = 2. Three category curves are shown in Figure 3.

Figure 3



The left-most curve in Figure 3 represents the probability function for all examinees scoring in the first category, i.e., getting a score of "0" (completely incorrect) on the item. Those of very low ability (e.g., -3 to -2) are very likely to be in this category, and in fact, are more likely to be in this category than the other two. Those scoring in the second category, getting a "1", tend to fall in the middle range of abilities (the middle curve.) The final, right-most curve represents the probability function for those examinees scoring in the third category, i.e., getting a score of "2" (completely correct). Very high ability examinees are clearly more likely to be in this category than in any other, but there are still some examinees of average and low abilities that can get full credit for the item.

Although the actual computations are somewhat complex, the points at which curves cross each other have a similar interpretation as for the dichotomous case. Consider the point at which the curve for the first category, score = 0, crosses the curve for the second category, score = 1. For abilities to the left of (or less than) this point, the probability is greatest for a category 1 response. To the right (or greater than) this point, and up to the point at which the curves cross for the second and third categories, scores 1 and 2, the most likely response is category 2, or score = 1. Note that the likelihood of a category 2 response declines in both directions as ability decreases to the low extreme or increases to the high extreme. These points then may be thought of as the difficulties of crossing the "steps" between categories.

Simultaneous calibration of items/tasks from different item types necessitates the use of a polytomous model that allows the number of score categories (typically score points on a scoring rubric) to vary across assessment modes. One of the popular polytomous models that can handle a mixing of item types is the Partial Credit Model (PCM).

The Rasch PCM is a direct extension of the dichotomous one-parameter IRT model developed by Rasch in the 1950s (Rasch, 1980). For an item/task involving m score categories, the general expression for the probability of scoring x on item/task i is given by

$$P_{ik}(x = k-1|\theta) = \frac{e^{\sum_{j=0}^{k-1}(\theta - B_{ij})}}{\sum_{k=1}^{m} e^{\sum_{j=0}^{k-1}(\theta - B_{ij})}}, \text{ where k=1,...,m (i.e., x=0,...,m-1)}$$

and by definition, $\sum_{j=0}^{0}(\theta - B_{i0}) = 0$

B_{ij} : the jth step difficulty parameter of item i

The equation gives the probability of an examinee scoring in a particular category (score = x) on an item/task, where i as a function of the person's position θ on the variable being measured and the step difficulties of the item/task. Specifically, the numerator involves only the particular category the examinee scored in and is equal to the logit (natural log-odds) of the sum of the differences between θ and B_{ij} for the completed steps associated with that category. The denominator is the sum of the numerator value for each category on the item.

With partial-credit modeling, the multiple-choice items and constructed-response questions are scaled in such a way that for each test form a single raw score to scaled score conversion is obtained. Such a scaling places item (and item-step) difficulties on the same scale as student abilities. This placement on a common scale allows the item difficulty of the multiple-choice items (or step value for the scores of 0 and 1) to be compared relative to the step (difficulty) values of the performance tasks. Note that the dichotomous one-parameter Rasch model is simply a special case of the Rasch PCM because dichotomous items can be treated as one-step items.

One important property of the Rasch PCM is the separability of estimation of item/task parameters and person parameters. Because of this separability property, the total score given by the sum of the categories in which a person responds is a sufficient statistic for estimating person ability (i.e., no additional information need to be obtained). Also, the total number of responses across examinees in a particular response category of an item is a sufficient statistic for estimating the step difficulty for that item's category.

Estimation equations are given in *Rating Scale Analysis* (Masters and Wright, 1982). The BIGSTEPS computer program was used to perform the Rasch Partial Credit Model analyses (Linacre and Wright, 1995).

II. Equating

New secured forms must continually be constructed for future test administrations. The test forms are equated so as to convert the raw scores obtained from two forms of the test so that the scores derived from the two forms *after conversion* will be directly equivalent. Different forms of the test are designed to have comparable item content and similar distributions of item statistics based on tryout administration. The equating adjusts for unintended differences in difficulty of the forms. Typically, and with the DSTP, equating adjusts raw test scores from different forms to a common scale so that identical scaled scores earned this year and last year reflect the same level of student achievement, even though the corresponding raw scores may differ.

IIa. Equating DSTP Reading and Mathematics

Equating of the DSTP reading and mathematics assessments was done with the Rasch PCM using the BIGSTEPS computer program and an anchor test design. The description of equating is based on the first two forms, 1998 and 1999, but applies to all future forms. Anchor items are the same items that appeared in both the 1998 Form and in the 1999 Form. For each assessment, about two-thirds of the items were in common between the two forms. The anchor items were used to develop a linking constant that places the item step values from the 1999 Form on the same logit scale as the 1998 Form. The linking constant is computed as the difference between the average step-value for the anchor items from the 1998 Form's BIGSTEPS analysis, minus the average step value from the 1999 Form's BIGSTEPS analysis. Adding this linking constant to the step values for each of the items in the 1999 Form places all of the 1999 Form's step values (and log ability estimates) on the same Rasch logit scale as the 1998 Form.

The DSTP reading assessment at each grade level was linked to the *Stanford Achievement Test Series, Ninth Edition* (SAT9) reading comprehension scaled scores. The DSTP mathematics assessment was linked to the SAT9 problem solving scaled scores at grades 3, 5, and 8, and to SAT9 mathematics at grade 10. (The SAT9 high school tests do not have a separate test of problem solving.) Linking to SAT9 follows the same procedure just described for equating two forms of the DSTP. The linking was accomplished by computing the SAT9 linking constant as the difference between the average step-value for the SAT9 items from the standardization, minus the average step value from the 1998 DSTP administration (i.e., the 1998 Form's BIGSTEPS analysis).

For each grade level of the DSTP reading assessment and mathematics assessment, the SAT9 linking constant plus the SAT9 standardization equating constant (developed by HEM to create the SAT9 scaled scores) are both added to the step-values for each of the items in the 1998 Form. In fact, the SAT9 linking constant and the SAT9 standardization

equating constant are added to the step-values of all items in future forms. For example, for the 1999 Form, the SAT9 linking constant, the SAT9 standardization equating constant, and the 1999 Form's linking constant (the linking constant that equates the 1999 Form to the 1998 Form) were all added to the step values from the 1999 Form's BIGSTEPS analysis.

Because the ability estimates are on a scale that includes negative and decimal values, the ability estimates were converted to a different metric through scaling. The scaling consisted of applying a linear transformation (multiplying by 40 and adding 400) to these Rasch log ability estimates. This linear transformation produced 3 digit, unit interval scaled scores that range from approximately 150 to 800 for each assessment (reading and mathematics) across grades (3, 5, 8, and 10).

Since both the multiplicative constant of 40 and the additive constant of 400 are different from the SAT9 linear transformation constants, the DSTP scaled scores are ***not*** the same as the SAT9 scaled scores. The DSTP scaled scores are only within a linear transformation of the SAT9 scaled scores. This was done to avoid misinterpretation or over-interpretation between scaled scores derived from two different tests. However, it was thought that linking the DSTP and the SAT9 could be of research interest.

IIb. Equating Science and Social Studies

Equating of the DSTP science and social studies assessments was also done with the Rasch PCM using the BIGSTEPS computer program and an anchor test design. The description of equating is based on the first two forms, 2000 and 2001, but applies to all future forms. For each assessment, about two-thirds of the items were in common between the two forms. The anchor items were used to develop a linking constant that places the item step values from the 2001 Form on the same logit scale as the 2000 Form. The linking constant is computed as the difference between the average step values for the anchor items from the 2000 Form's BIGSTEPS analysis, minus the average step value from the 2001 Form's BIGSTEPS analysis. Adding this linking constant to the step values for each of the items in the 2001 Form places all of the 2001 Form's step values (and log ability estimates) on the same Rasch logit scale as the 2000 Form.

Since the SAT9 Science and Social Studies tests are ***not*** administered with the DSTP science and social studies assessments, there can not be any linking to the SAT9 as was done with the DSTP reading and mathematics assessments. Therefore, there is no SAT9 linking constant and SAT9 standardization equating constant, but only the 2001 linking constant. Of course, with each new year/form, there is another linking constant relating that new year to the previous year's form.

Again, because the ability estimates are on a scale that includes negative and decimal values, the ability estimates were converted to a different metric through scaling. This time the linear transformation of these Rasch log ability estimates was tied to the standard setting for each of these assessments. (For a complete description, see the *Report and Recommendations to the Delaware State Board of Education for: Establishing*

Proficiency Levels for the Delaware Student Testing Program in Science and Social studies for Grades 8 & 11 and grades 4 and 6). The standard setting for each assessment established four cut-scores yielding the five Performance Levels: Distinguished, Exceeds the Standard, Meets the Standard, Below the Standard, and Well below the Standard. (A cut-score is the lowest score corresponding to a performance level.) The linear scaling of the Rasch log ability estimates for each assessment was done so that the cut-score for Meets the Standard was set to 300 and the cut score for Exceeds the Standard was set to 325. Thus, all scaled scores between 300 and 324 yield Meets the Standard and a scaled score of 325 is the lowest scaled score for Exceeds the Standard. This is true for each of the grades 4, 6, 8, and 11 science and social studies assessments, and is true for all forms of each assessment; 2000, 2001 and all future forms. The lowest scaled score for Distinguished and for Below the Standard cannot be fixed across these assessments and across the different forms because only two cut scores can be set to predetermined values with a linear transformation. The resulting linear transformation produced 3 digit, unit interval scaled scores that can range from approximately 150 to 450 (the actual range varies by grade, assessment, and form).

III. 2001 Equating and Scaling Results

The ultimate results of equating and scaling are, of course, the Raw Score to Scaled Score Conversion Tables presented in Appendix I for all assessments.

Recall that equating involves comparing the step values for the anchor items from the 2001 Form with those from the 2000 Form. Appendix J contains the plot of the 2001 step values versus the 2000 step values for the anchor items for each assessment. The number of plotted points for an assessment ranges from 41 for grade 6 science to 61 for grade 3 reading. Also shown in each plot is the 45-degree straight line that passes through the mean of the 2001 step values and the mean of the 2000 step values. The plots show that the step values fall along this 45-degree line as the model requires. Of course, not all points are on or right next to the line due to the inherent error that is in all measurement, and occasionally, a point is quite far from the line. Across the 16 assessments, grade 8 science shows the greatest dispersion of points from the line with three points that are quite far from it. In fact, these three points are among the four points that are the furthest from the line in all 16 plots. Another way to evaluate the plots is to compute the correlation coefficient between the 2001 step values and the 2000 step values. The maximum value for the coefficient is 1.00. A value of .00 indicates no linear relationship between the step values from 2001 and 2000. The correlation coefficient (r) is given in the upper right-hand corner of each plot. Across all the 2001 assessments, the correlations range from .914 to .994. The correlation of .914 is for grade 8 science, but the next lowest correlation is .966. Thus, 15 of the 16 correlations range from .966 to .994, which are values as close to 1.00 as can practically be expected. But even the lowest correlation is over .900.

Part Five. Technical Characteristics of the DSTP

This section focuses on the technical characteristics of the 2001 DSTP reading, mathematics, writing, science, and social studies. Statistics are presented on the reliability of test scores, standard error of measurement, rater consistency, and correlation matrix.

I. Reliability for the DSTP

Test reliability refers to the accuracy of scores. Tests with high reliabilities provide scores that are stable over time and between test forms. Reliability is a necessary condition for good quality assessment, and it is important to establish test reliabilities through empirical studies so that sound judgments can be made. The reliability of a test is a function of the test content, length of the test, item difficulty, the standard deviation, and the procedure for test development, test administration, and other factors. The standard error of measurement provides an indicator of the accuracy of the test scores using the observed score scale. The magnitude of standard error of measurement depends on the standard deviation and the reliability of the tests.

Cronbach's alpha

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \text{var}(Y_i)}{\text{var}(Y_{tot})}\right)$$

where
k = number of items on the test
var (Y_i) = variance of item i
var (Y_{tot}) = total test variance

Standard Error

$$SEM = SD(Y_{tot})\sqrt{1 - reliability}$$

where
SD (Y_{tot}) = Standard deviation of the test

Ia. Estimate of Reliabilities and Standard Error of Measurement

Table 4 presents reliability coefficients (Cronback's alpha) and standard errors of measurement based on the scale scores in reading, mathematics, science, and social studies by grade and test in the 2001 DSTP administration. The reliability coefficients are ranging from .91 to .92 in reading and mathematics, .88 to .91 in science, and .90 to .93 in social studies across grades. The values of standard error of measurement range

from 11.2 to 11.9 in reading, 10.6 to 12.7 in mathematics, 5.2 to 8.5 in science, and 5.3 to 8.1 in social studies across grades.

Ib. Correlations

Tables 5 and 6 present the correlation coefficients between SAT9 reading comprehension and Delaware-developed items in reading, SAT9 mathematics problem solving for grades 3, 5, and 8 or SAT9 mathematics for grade 10 and Delaware-developed items in mathematics, and among the four content areas in science and social studies. Correlations among different item formats of a given test are also calculated. The results show that the correlation coefficients between SAT9 and Delaware-developed items range from .72 to .76 in reading and .76 to .82 in mathematics across grades. The correlations between SAT9 reading and Delaware-developed constructed-response items range from .57 to .66 in reading and .73 to .76 in mathematics across grades.

Table 7 presents the correlation matrix among the total writing scores, stand-alone writing scores, text-based writing scores, reading standard-based scores, and SAT9 reading scores. The correlations between stand-alone and text-based writing scores are moderately low, .46 in grade 3, .52 in grade 5, .46 in grade 8, and .43 in grade 10. The data show that writing scores are moderately associated with reading scores, ranging from .53 to .61, and with SAT9 reading scores, ranging from .52 to .56 across grades.

Table 8 presents the inter-correlation coefficients among the four sub-content areas, inquiry, physical science, earth science, and life science, in science. Moderate to moderately high correlations are observed among the sub-content areas, ranging from .60 to .72 across grades. The correlation between multiple-choice and short answer items is .73 in grade 4, .71 in grade 6, .76 in grade 8, and .76 in grade 11. Table 9 presents the inter-correlation coefficients among the four sub-content areas, civics, economics, geography, and history, in social studies. Moderately high correlations are observed among the sub-content areas, ranging from .64 to .79 across grades. The correlation between multiple-choice and short answer items is .67 in grade 4, .69 in grade 6, .73 in grade 8, and .69 in grade 11.

Ic. Rater Consistency

Students' responses to the stand-alone writing prompt were evaluated by two raters using the holistic scoring rubric on a 5-point scale. The sum of the two scores were reported as the stand-alone writing score. Table 10 shows the correlations between the writing scores on the stand–alone writing from the two raters: the rater's correlations, the percentage of perfect agreement, and the percentage of plus/minus one-point agreement. Moderate correlation coefficients between the two raters' scores are shown, .58 in grade 3, .67 in grade 5, .57 in grade 8, and .61 in grade 10. The average percentage of perfect agreement is ranging from 66.4 in grade 5 to 71.1 in grade 8. The percentage of agreement between the two raters within one score point is ranging from 98.9 in grade 10 to 99.5 in grade 8.

One rater was used to score the text-based writing and all constructed-response items (including short-answer and extended construct-response items) in reading, mathematics, science, and social studies in 2001. The Performance Assessment Scoring Center (PASC) at Harcourt Educational Measurement established a system of Rater Monitoring and Quality Assurance Checks to ensure continuing quality and refinement of scoring after training. The ongoing process included:

- Read-Behind: Team Leader and/or Room Director spot checks individual raters' scored papers. Typically, these master raters should review about 10% of student papers in the initial stages of the project. In the later stage, the leader might randomly check different raters scoring.

- Calibrations: Pre-selected sets of 5 papers or items with clear-cut score points were distributed to raters on a daily basis for the purpose of keeping them on the track and preventing rater 'drift'. Three out of 5 scores must be perfect matches to the set scores to be acceptable and other scores must be adjacent. If a rater fell below the 60% on two consecutive days, PASC required a formal re-training with accompanying documentation.

- Resolution: Non-adjacent scores required a third "resolution" reading. The percentage of papers requiring resolution was closely monitored for every holistic scoring rater.

- Monitoring Reports: Daily reports and Cumulative reports were prepared during scoring about the project and rater performance.

II. Item and Test Statistics

Table 11 shows the means and standard deviations of item difficulty and item discrimination (point-biserial correlation) for all core items by grade and test. Using the classical approach, relative mean was calculated for short the relative mean for answer and extended constructed-response questions. The average p-values are .59, .63, .62, and .61 for grades 3, 5, 8, and 10, respectively, in reading; .67, .54, .49, and .41 for grade 3, 5, 8, and 10, respectively, in mathematics; .65, .52, .41, and .40 for grades 4, 6, 8, and 11, respectively, in science; and .53, .48, .43, and .34 for grades 4, 6, 8, and 11, respectively, in social studies. The average point-biserial correlations are .37, .40, .35, and .39 for grades 3, 5, 8, and 10, respectively, in reading; .39, .39, .37, and .37 for grades 3, 5, 8, and 10, respectively, in mathematics; .37, .37, .38, and .40 for grades 4, 6, 8, and 11, respectively in science; and .40, .43, .44, and .46 for grades 4, 6, 8, and 11, respectively, in social studies. The histogram distributions of item difficulties by grade and test can be found in Appendix K and the histogram distributions of point-biserial correlations by grade and test are in Appendix L.

Table 1
Statistics of Reading, Writing, and Mathematics Scores

Grade	Statistics	Reading		Writing			Mathematics	
		SAT9	SBS	Total	Text-based	Stand-alone	SAT9	SBS
3	N	8852	8853	8810	8680	8851	8850	8851
	Mean	56.20	433.11	5.87	1.65	4.23	59.80	429.22
	S.D.	19.00	39.56	1.70	0.78	1.20	21.70	43.32
5	N	8520	8520	8539	8406	8522	8521	8518
	Mean	52.10	466.14	7.30	2.11	5.23	55.80	459.27
	S.D.	20.10	44.23	1.97	0.75	1.42	23.30	41.49
8	N	8701	8695	8729	8506	8688	8664	8657
	Mean	54.30	510.20	7.87	2.43	5.53	50.80	484.30
	S.D.	21.10	39.60	1.73	0.77	1.10	20.80	40.39
10	N	8001	7995	8062	7703	7987	7933	7929
	Mean	46.40	507.17	7.33	2.01	5.46	49.90	514.07
	S.D.	19.00	42.31	1.85	0.59	1.34	21.70	38.13

SBS - Standar-based score

Table 2
Statistics of Science and Social Studies Sub-Scores

Grade	Statistics	Science				Social Studies			
		Inquiry	Physaical Science	Earth Science	Life Science	Civics	Economics	Geography	History
4	N	8910	8910	8910	8910	8905	8905	8905	8905
	Mean	13.05	9.10	7.21	14.36	8.95	9.96	6.84	9.83
	S.D.	3.56	2.63	2.36	3.41	3.20	3.22	3.09	3.19
6	N	8978	8978	8978	8978	8953	8953	8953	8953
	Mean	6.70	11.08	8.37	8.83	8.39	8.43	7.08	7.32
	S.D.	2.72	3.09	2.94	3.59	3.18	3.39	3.31	3.55
8	N	8561	8561	8561	8561	8542	8542	8542	8542
	Mean	5.21	6.01	6.93	9.40	6.47	7.10	7.87	7.25
	S.D.	2.70	3.11	3.01	4.17	3.41	3.51	3.92	3.63
11	N	6202	6202	6202	6202	6108	6108	6108	6108
	Mean	4.72	7.61	6.87	7.42	5.80	5.33	5.97	5.71
	S.D.	2.51	3.57	2.98	4.23	3.39	3.56	3.45	3.50

Table 3
Percentage of Students in Each Performance Level by Grade and Test

Grade	Level	Reading		Writing		Mathematics		Grade	Science		Social Studies	
		N.	%	N.	%	N.	%		N.	%	N.	%
3	Well Below	1170	13.2	1651	18.7	1216	13.7	4	319	3.6	1408	15.8
	Below	1301	14.7	4299	48.8	1390	15.7		910	10.2	2588	29.1
	Meets	4407	49.8	2823	32.0	4335	49.0		5193	58.3	4133	46.4
	Exceeds	1102	12.4	34	0.4	1384	15.6		1738	19.5	547	6.1
	Distinguished	873	9.9	3	0.0	526	5.9		750	8.4	229	2.6
	Total	8853	100.0	8810	100.0	8851	100.0		8910	100.0	8905	100.0
5	Well Below	1562	18.3	1509	17.7	1714	20.1	6	767	8.5	1844	20.6
	Below	1476	17.3	2742	32.1	1565	18.4		1949	21.7	2413	27.0
	Meets	4008	47.0	3940	46.1	4018	47.2		4986	55.5	3977	44.4
	Exceeds	822	9.6	327	3.8	810	9.5		986	11.0	498	5.6
	Distinguished	652	7.7	21	0.2	411	4.8		290	3.2	221	2.5
	Total	8520	100.0	8539	100.0	8518	100.0		8978	100.0	8953	100.0
8	Well Below	1523	17.5	753	8.6	3056	35.3	8	2283	26.7	2226	26.1
	Below	1609	18.5	2199	25.2	2174	25.1		2636	30.8	2242	26.2
	Meets	4963	57.1	5440	62.3	2293	26.5		2735	31.9	2863	33.5
	Exceeds	410	4.7	331	3.8	568	6.6		614	7.2	707	8.3
	Distinguished	190	2.2	6	0.1	566	6.5		293	3.4	504	5.9
	Total	8695	100.0	8729	100.0	8657	100.0		8561	100.0	8542	100.0
10	Well Below	1800	22.5	1153	14.3	2948	37.2	11	1538	24.8	2250	36.8
	Below	1547	19.3	2375	29.5	2258	28.5		1869	30.1	1844	30.2
	Meets	4264	53.3	4357	54.0	1762	22.2		2260	36.4	1381	22.6
	Exceeds	306	3.8	173	2.1	371	4.7		333	5.4	321	5.3
	Distinguished	78	1.0	4	0.0	590	7.4		202	3.3	312	5.1
	Total	7995	100.0	8062	100.0	7929	100.0		6202	100.0	6108	100.0

Table 4
Reliability Coefficients of Test Scores by Grade and Test

Grade	2001							
	Reading		Math		Science		Social Studies	
	Reliability	SEM	Reliability	SEM	Reliability	SEM	Reliability	SEM
3	0.91	11.40	0.91	12.70				
4					0.88	5.20	0.90	5.80
5	0.92	11.90	0.92	11.30				
6					0.88	5.60	0.91	5.30
8	0.91	11.20	0.92	11.00	0.89	8.00	0.92	6.80
10	0.92	11.40	0.92	10.60				
11					0.91	6.50	0.93	8.10

SEM: Standard error of measurement. The SEM was calculated using scale scores.

Table 5
Correlation Matrix in Reading

Grade 3	SAT 9 Reading Comprehension	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.74	0.61	0.58	0.66	0.76
DEL MC		1.00	0.65	0.61	0.70	*
DEL SA			1.00	0.66	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

Grade 5	SAT 9 Reading Comprehension	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.75	0.61	0.55	0.64	0.75
DEL MC		1.00	0.70	0.61	0.72	*
DEL SA			1.00	0.66	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

Grade 8	SAT 9 Reading Comprehension	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.74	0.52	0.49	0.57	0.72
DEL MC		1.00	0.59	0.54	0.64	*
DEL SA			1.00	0.60	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

Grade 10	SAT 9 Reading Comprehension	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.72	0.62	0.55	0.63	0.72
DEL MC		1.00	0.70	0.66	0.73	*
DEL SA			1.00	0.74	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

* Due to confound effect, the correlation coefficient is not shown.
MC = Multiple-choice item
SA = Short answer item
ECR = Extended constructed-response item
CR = Constructed-response items, including short answer and extended constructed-response items

Table 6
Correlation Matrix in Mathematics

Grade 3	SAT 9 Math Problem Solving	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.78	0.74	0.63	0.76	0.82
DEL MC		1.00	0.73	0.60	0.75	*
DEL SA			1.00	0.62	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

Grade 5	SAT 9 Math Problem Solving	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.76	0.72	0.65	0.75	0.79
DEL MC		1.00	0.75	0.69	0.78	*
DEL SA			1.00	0.71	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

Grade 8	SAT 9 Math Problem Solving	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.73	0.69	0.66	0.73	0.78
DEL MC		1.00	0.72	0.69	0.76	*
DEL SA			1.00	0.71	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

Grade 10	SAT 9 Mathematics	Delaware-Developed Items				Delaware Total
		MC	SA	ECR	CR	
SAT Reading	1.00	0.68	0.72	0.61	0.73	0.76
DEL MC		1.00	0.72	0.63	0.74	*
DEL SA			1.00	0.70	*	*
DEL ECR				1.00	*	*
DEL CR					1.00	*
DEL Total						1.00

* Due to confound effect, the correlation coefficient is not shown.
MC = Multiple-choice item
SA = Short answer item
ECR = Extended constructed-response item
CR = Constructed-response items, including short answer and extended constructed-response items

Table 7
Correlation Matrix in Reading and Writing

	Writing Scores			Reading Scores	
Grade 3	**Writing Total**	**Text-Based**	**Stand-Alone**	**Reading Scale**	**SAT9 Reading**
Writing Total	1.00	*	*	0.61	0.55
Text-based		1.00	0.46	0.57	0.46
Stand-Alone			1.00	0.50	0.47
Reading Scale				1.00	*
SAT9 Reading					1.00
Grade 5					
Writing Total	1.00	*	*	0.59	0.52
Text-based		1.00	0.52	0.53	0.42
Stand-Alone			1.00	0.52	0.47
Reading Scale				1.00	*
SAT9 Reading					1.00
Grade 8					
Writing Total	1.00	*	*	0.56	0.53
Text-based		1.00	0.46	0.47	0.37
Stand-Alone			1.00	0.48	0.50
Reading Scale				1.00	*
SAT9 Reading					1.00
Grade 10					
Writing Total	1.00	*	*	0.53	0.56
Text-based		1.00	0.43	0.47	0.39
Stand-Alone			1.00	0.43	0.52
Reading Scale				1.00	*
SAT9 Reading					1.00

* Due to confound effect, the correlation coefficient is not shown.

Table 8
Correlation Matrix in Science

Grade 4	Sub-Content Areas				Item Format	
	Inquiry	Physical S.	Earth S.	Life S.	MC	SA
Inquiry	1.00	0.64	0.60	0.66		
Physical Science		1.00	0.58	0.63		
Earth Science			1.00	0.61		
Life Science				1.00		
Multiple-choice					1.00	0.73
Short Answer						1.00

Grade 6	Sub-Content Areas				Item Format	
	Inquiry	Physical S.	Earth S.	Life S.	MC	SA
Inquiry	1.00	0.60	0.60	0.64		
Physical Science		1.00	0.60	0.65		
Earth Science			1.00	0.64		
Life Science				1.00		
Multiple-choice					1.00	0.71
Short Answer						1.00

Grade 8	Sub-Content Areas				Item Format	
	Inquiry	Physical S.	Earth S.	Life S.	MC	SA
Inquiry	1.00	0.64	0.65	0.67		
Physical Science		1.00	0.64	0.65		
Earth Science			1.00	0.70		
Life Science				1.00		
Multiple-choice					1.00	0.75
Short Answer						1.00

Grade 11	Sub-Content Areas				Item Format	
	Inquiry	Physical S.	Earth S.	Life S.	MC	SA
Inquiry	1.00	0.70	0.65	0.72		
Physical Science		1.00	0.69	0.72		
Earth Science			1.00	0.70		
Life Science				1.00		
Multiple-choice					1.00	0.76
Short Answer						1.00

MC = Multiple-choice item
SA = Short answer item

Table 9
Correlation Matrix in Social Studies

Grade 4	Sub-content Areas				Item Format	
	Civics	Economics	Geography	History	MC	SA
Civics	1.00	0.71	0.64	0.69		
Economics		1.00	0.67	0.71		
Geography			1.00	0.67		
History				1.00		
Multiple-choice					1.00	0.67
Short Answer						1.00

Grade 6	Sub-content Areas				Item Format	
	Civics	Economics	Geography	History	MC	SA
Civics	1.00	0.72	0.69	0.73		
Economics		1.00	0.71	0.74		
Geography			1.00	0.73		
History				1.00		
Multiple-choice					1.00	0.69
Short Answer						1.00

Grade 8	Sub-content Areas				Item Format	
	Civics	Economics	Geography	History	MC	SA
Civics	1.00	0.73	0.76	0.76		
Economics		1.00	0.76	0.74		
Geography			1.00	0.78		
History				1.00		
Multiple-choice					1.00	0.73
Short Answer						1.00

Grade 11	Sub-content Areas				Item Format	
	Civics	Economics	Geography	History	MC	SA
Civics	1.00	0.79	0.78	0.77		
Economics		1.00	0.79	0.79		
Geography			1.00	0.79		
History				1.00		
Multiple-choice					1.00	0.69
Short Answer						1.00

MC = Multiple-choice item
SA = Short answer item

Table 10
Raters' Correlation and Raters' Agreement for Writing in 2001

	Grade 3	Grade 5	Grade 8	Grade 10
Rater's Correlation	0.58	0.67	0.57	0.61
Perfect Agreement (%)	68.4	66.4	71.1	65.8
Plus/Minus 1-point Agreement (%)	99.4	99.3	99.5	98.9

Table 11
Summary of Item Statistics by Test and Grade

Reading Grade	Difficulty Mean	S. D.	Grade	Point Biserial Mean	S. D.
3	0.59	0.18	3	0.37	0.09
5	0.63	0.16	5	0.40	0.08
8	0.62	0.16	8	0.35	0.09
10	0.61	0.17	10	0.39	0.11
Mathematics Grade					
3	0.67	0.18	3	0.39	0.11
5	0.54	0.18	5	0.39	0.11
8	0.49	0.18	8	0.37	0.11
10	0.41	0.17	10	0.37	0.11
Science Grade					
4	0.65	0.17	4	0.37	0.09
6	0.52	0.21	6	0.37	0.12
8	0.41	0.17	8	0.38	0.13
11	0.40	0.21	11	0.40	0.11
Social Studies Grade					
4	0.53	0.22	4	0.40	0.10
6	0.4(	0.18	6	0.43	0.12
8	0.43	0.19	8	0.44	0.14
11	0.34	0.20	11	0.46	0.13

Part Six. References

Delaware Student Testing Program: State Summary Report – 2001 Administration

Delaware Student Testing Program: 2001 State Report of Student Questionnaire Survey

Delaware Student Testing Program: 2001 State Summary of Writing Report

Delaware Student Testing Program: Technical Report 1998-2000

Oh, H-J., Nandakumar, R., Glutting, J. & Zhang, L. DIF in Mathematics on Delaware Student Testing Program for 3rd Grade: Comparison of Mantel-Haenszel, Logistic Regression and SIBTEST Procedures. Paper presented at the 2002 AERA Annual Conference, April, New Orleans, LA.

Holland, P. W. & Wainer, H. (Eds.), Differential Item Functioning (pp. 3-23). Hillsdale. NJ: Erlbaum.

Nandakumar, R. (1993). Simultaneous DIF amplication and cancellation: Shealy-Stout's test for DIF. Journal of Educational Measurement, 30, 4, 293-311.

Report and recommendations to the Delaware State Board of Education for: Establishing proficiency levels for the Delaware Student Testing Program in science and Social studies – Grades 8 & 11, September 20, 2001.

Report and recommendations to the Delaware State Board of Education for: Establishing proficiency levels for the Delaware Student Testing Program in science and Social studies – Grades 4 & 6, February 21, 2002.

Shealy, R. & Strout, W.F. (1993a). An item response theory model for test bias. In P.W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 197-239). Hillsdale. NJ: Erlbaum.

Shealy, R. & Strout, W.F. (1993b). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as Item bias/DIF. Psychometrika, 58, 159-194.

Standards for Educational and Psychological Testing, 1999

State of Delaware – English Language Arts Curriculum Framework

State of Delaware – Mathematics Curriculum Framework

State of Delaware – Science Curriculum Framework

State of Delaware – Social Studies Curriculum Framework

Appendix A

Test Development Committees

Test Development Committee for English Language Arts

Name	Affiliation	Year(s) of Service
Deidra Aikens	Christina District, Maclary Elementary	2001
Bonnie Albertson	University of Delaware Delaware Reading/Writing Project	1996 - 2002
Darlene Bolig	Department of Education	1996 - 2002
Mike Boyd	Lake Forest School District Lake Forest High School	2000 - 2002
Dawn Downes	Christina School District Eden Support Services Center	2000 - 2002
Christine Evans	University of Delaware Delaware Reading/Writing Project	1996 - 2002
Marty Hodgkins	Appoquinimink School District Reading Intermediate School	1996 - 2002
Mike Kelley	Department of Education	1996 - 2002
Lorelei Meanor	Department of Education	2000 - 2002
Deanne McCredie	Cape Henlopen School District Milton Middle School	1998 - 2002
Jane Ragins	Capital School District William Henry Middle School	2001 - 2002
Jacklyn Shockley	Cape Henlopen School District Shields Elementary School	2000 - 2002
Kate Szegda	Newark Charter Newark Charter School	1998 - 2002
Aleta Thompson	Cape Henlopen School District Cape Henlopen High	2001 - 2002
Carol Vukelich	University of Delaware Delaware Center for Teacher Education	1996 - 2002
Denise Weiner	Brandywine School District Springer Middle School	2001 – 2002

Test Development Committee for Math

Name	Affiliation	Year(s) of Service
Sally Caldwell	Department of Education	1998 - 2002
Susan Carlin	Christina School District	1998 - 2002
Maureen Leclerc	Cape Henlopen School District	1998 - 2002
John Matthias	Red Clay School District	1998 - 2002
Valerie Maxwell	Appoquinimink School District	1998 - 2002
Susan Nancarrow	Seaford School District	1999 - 2002
Jan Parsons	Indian River School District	1998 – 2002
Jan Shetzler	Polytech School District	1998 - 2002
Carol Stead	New Castle County Vo Tech School District	1998 - 2002
Mary Lynn Vincent	Colonial School District	1998 – 2002
Shirley Ellison	Red Clay School District	2002
Vicky Pendleton	Indian River School District	2002
Wendy Harrington	Cape Henlopen School District	2002

Test Development Committee for Science

Name	Affiliation	Year(s) of Service
John Berry	Lake Forest School District	2001-2002
Mary Bing	Laurel School District	1998-2000
Henry E. Bouchelle	Colonial School District	1998-2000
Jane H. Carey	Brandywine School District	1998-2000
Julie A. Hanenfeld	Seaford School District	1999-2000
Paula S. Henderson	Christina School District	1998-2000
Michelle Kutch	Brandywine School District	1999-2000
Faye Markowitz	Christina School District	2000-2001
Tonyea Mead	Milford School District	1999-2001
Eugene E. Montano	Capital School District	2001
Carolyn C. Newsom	Brandywine School District	1998-2001
Carole M. Palmer	Cape Henlopen School District	1998-2001
Randall J. Redard	Cape Henlopen School District	1998-2001
Julie A. Schmidt	University of Delaware	1995-2001
Tom Shaffer	Sussex Technical School District	1998-2001
Gwyneth Sharp	Cape Henlopen School District	1998-2001
Janice Trainer	Christina School District	1998-2001
Linda Willey-Impagliazzo	Christina School District	1998-2001
Sandra K. Wolford	Colonial School District	1999-2001

Science Advisory Committee

Name	Affiliation	Year(s) of Service
Barbara J. Duch	University of Delaware	1998-2001
Kelli Martin	Appoquinimink School District	2001
William J. McIntosh	Delaware State University	1998-2001
Amy D. Quillen	Smyrna School District	2001

Test Development Committee for Social Studies

Name	Affiliation	Year(s) of Service
Dr. Anthony M. Armstrong	Wesley College Political Science Department	1999 - 2002
Michael Brelick	New Castle County Vo-Tech Delcastle Technical High School	1999 - 2001
Hilton Cohen	Christina District Retired	2000 - 2002
John Crum	Brandywine District Mt. Pleasant High School	1998 - 2002
Barbara Emery	Christina District Retired	2000 - 2002
Sue George	Caesar Rodney District Welch Elementary	2001 - 2002
Mary K. Hall	Christina District Glasgow High School	2001 - 2001
Charlotte Hughes	Red Clay District Retired	1998 - 2002
Robert B. Maull, Jr.	Seaford District Seaford Middle	2001 - 2002
Bonnie Meszaros	University of Delaware Center for Economic Education	1998 - 2002
James B. O'Neill	University of Delaware Center for Economic Education	1998 - 2002
Gerald Peden	Cape Henlopen District Cape Henlopen High School	1999 - 2001
Joann F. Prewitt	Department of Education	1998 - 2002

Test Development Committee for Social Studies

Name	Affiliation	Year(s) of Service
Judy Purcell	Milford District Banneker Elementary	1998 - 2002
Rebecca Reed	Colonial District Gunning Bedford Middle	1998 - 2002
Peter Rees	University of Delaware Department of Geography	1998 - 2002
Preston W. Shockley, III	Cape Henlopen District Lewes Middle School	2001 - 2002
Dawn Willis	Milford District Milford Middle School	1998 - 2002

Appendix B

Bias Review Committee

Bias Review Committee

Name	Affiliation	Year(s) of Service
Mark Abbott	Sussex Tech Sussex Tech High School	1998-2002
Wendy Balakhani	Christina School District Sterck School for the Deaf	1999-2002
Ariadna Clare	Red Clay School District Administrative Office	1998-2002
Shirley Connoway	Lake Forest School District Lake Forest East Elementary	1998-2002
Kathy Cuputo	Christina School District Sterck School for the Deaf	2001-2002
Charlene Dolgos	DE Health & Social Services Division for the Visually Impaired	1999-2002
Judy Goldbaum	Brandywine School District Hanby Middle School	1999-2002
Phyllis Heimall	Caesar Rodney School District Welch Elementary	1998-2002
Mike Kijowski	Caesar Rodney School District Fifer Middle School	2002
Rebecca Lykens	Retired	1998-2002
Terrance Moore	Woodbridge School District Woodbridge Middle School	2000-2002
Gwyneth Sharp	Department of Education Science Resource Center	1998-2002
George Smith	Community City Official	2002
Cathy Williams	Retired	1999-2002
Colleen Wozniak	Department of Education Unified Planning & Quality Assurance	1998-2002

Appendix C

Technical Advisory Committee

DSTP Technical Advisory Committee

Name	Affiliation	Year(s) of Service
Dr. Robert Calfee	University of California-Riverside	1998 – 2002
Dr. Steve Dunbar	University of Iowa	1996 – 2002
Dr. Ronald Hambleton	University of Massachusetts	2000 – 2002
Dr. Suzanne Lane	University of Pittsburgh	2000 - 2002
Dr. Ken Olsen	Mid-South Regional Resource Ctr.	1996 – 2002
Dr. Martha Thurlow	National Center on Educational Outcomes, University of Minnesota	1996 – 2002

Appendix D

Test Specifications for Reading

Test Specifications for Reading
by Passage Type

Grade	Literary		Informative		Technical		Total	
	%	*Max. Points*	%	*Max. Points*	%	*Max. Points*	%	*Max. Points*
3	65	53	25	20	10	8	100	81
5	60	50	30	25	10	9	100	84
8	55	45	25	20	20	16	100	81
10	50	42	25	21	25	21	100	84

Test Specifications for Reading
by Stance

Grade	Determining Meaning		Interpreting Meaning		Extending Meaning		Total	
	%	*Max. Points*	%	*Max. Points*	%	*Max. Points*	%	*Max. Points*
3	45	37	35	28	20	16	100	81
5	35	29	45	38	20	17	100	84
8	30	24	45	37	25	20	100	81
10	25	21	45	38	30	25	100	84

* The percentage and maximum points of each cell is approximation. The actual numbers may vary from year to year.

Appendix E

Test Specifications for Mathematics

Test Specifications for Mathematics
by Content Standards

Content Standards	Grade 3		Grade 5		Grade 8		Grade 10	
	N. of Items	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*
Estimation, Measurement & Computation	15	17	15	17	14	16	18	19
Number Sense	13	18	12	18	11	14	4	7
Algebra	5	6	5	6	12	14	14	15
Spatial Sense & Geometry	10	14	9	11	6	8	7	8
Statistics & Probability	8	11	10	14	11	15	12	17
Patterns, Relationship & Functions	9	11	9	11	7	11	6	12
Total	**60**	**77**	**60**	**77**	**61**	**78**	**61**	**78**

* The number of items and maximum points of each cell is approximation. The actual numbers may vary from year to year.

Test Specifications for Mathematics
by Cognitive Level

Cognitive Level	Grade 3		Grade 5		Grade 8		Grade 10	
	N. of Items	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*
Conceptual Knowledge	24	30	24	29	25	29	25	31
Procedual Knowledge	28	32	28	32	26	28	28	29
Mathematical Process	8	15	8	16	10	21	8	18
Total	**60**	**77**	**60**	**77**	**61**	**78**	**61**	**78**

* The number of items and maximum points of each cell is approximation. The actual numbers may vary from year to year.

Test Specifications for Mathematics
by Content Standards for 2001*

Content Standards	Grade 3		Grade 5		Grade 8		Grade 10	
	N. of Items	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*
Estimation, Measurement & Computation	16	19	19	27	9	9	19	22
Number Sense	12	16	9	9	14	19	3	4
Algebra	5	7	8	10	11	12	12	15
Spatial Sense & Geometry	11	15	8	10	9	12	7	8
Statistics & Probability	8	11	10	14	11	15	13	18
Patterns, Relationship & Functions	8	9	6	7	7	11	7	11
Total	**60**	**77**	**60**	**77**	**61**	**78**	**61**	**78**

Test Specifications for Mathematics
by Cognitive Level for 2001*

Cognitive Level	Grade 3		Grade 5		Grade 8		Grade 10	
	N. of Items	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*
Conceptual Knowledge	25	28	24	28	29	35	22	26
Procedual Knowledge	25	28	26	27	24	28	31	33
Mathematical Process	10	21	10	22	8	15	8	19
Total	**60**	**77**	**60**	**77**	**61**	**78**	**61**	**78**

* The discrepancy between the original test specifications and the one for 2001 due to the following reasons:
1) There is great overlap between Standard 5 Estimation, Measurement, and Computation and Standard 6 Number Sense;
2) The item pool was short of certain type of items; and
3) The DSTP mathematics test includes short answer (2-points) and extended constructed-response (4-points) items. Thus, the discrepancy of number of items of each cell from the original test specifications might result in greater difference of the maximum points.

Appendix F

Test Specifications for Science

Test Specifications for Science*

Grade	Inquiry		Life Science		Earth Science		Physical Science		Total
	N. of Items	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*Item (Point)*
4	16	22	16	22	9	12	9	12	**50 (68)**
6	13	18	13	18	12	16	12	16	**50 (68)**
8	8	11	16	22	12	16	14	19	**50 (68)**
11	8	11	16	22	12	16	14	19	**50 (68)**

* The Test Specifications for science shown here is the general one. The actual numbers may vary from year to year.

Test Specifications for Science*
for 2001

Grade	Inquiry		Life Science		Earth Science		Physical Science		Total
	N. of Items	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*Item (Point)*
4	15	20	16	22	9	12	10	14	**50 (68)**
6	10	12	14	20	13	18	13	18	**50 (68)**
8	9	12	16	22	12	17	13	17	**50 (68)**
11	9	12	15	21	12	16	14	19	**50 (68)**

* The discrepancy of the number of items and the maximum points for some sub-content areas due the short of item pool.

Appendix G

Test Specifications for Social Studies

Test Specifications for Social Studies

Grade	Civics		Economics		Geography		History		Total
	N. of Items	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*N. of Items*	*Max. Points*	*Item (Point)*
4	12	17	12	17	12	17	12	17	**48 (68)**
6	12	17	12	17	12	17	12	17	**48 (68)**
8	12	17	12	17	12	17	12	17	**48 (68)**
11	12	17	12	17	12	17	12	17	**48 (68)**

* The number of items and maximum points of each cell is approximation. The actual numbers may vary from year to year.

Appendix H

Frequency Distributions of Test Scores

Frequency Distributions of Scale Scores in Reading

Grade 3			Grade 5			Grade 8			Grade 10		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
194	1	0.01	288	1	0.01	343	1	0.01	338	1	0.01
273	1	0.01	300	1	0.01	353	1	0.01	351	3	0.04
283	1	0.01	319	1	0.01	369	3	0.03	360	3	0.04
291	1	0.01	326	1	0.01	375	4	0.05	369	3	0.04
299	6	0.07	333	6	0.07	381	4	0.05	376	6	0.08
306	7	0.08	338	7	0.08	386	6	0.07	382	9	0.11
312	10	0.11	344	9	0.11	391	7	0.08	388	11	0.14
317	8	0.09	349	20	0.23	396	8	0.09	393	8	0.10
322	17	0.19	353	16	0.19	400	7	0.08	398	29	0.36
327	13	0.15	358	17	0.20	404	31	0.36	402	28	0.35
332	16	0.18	362	22	0.26	408	12	0.14	406	28	0.35
336	22	0.25	366	25	0.29	412	17	0.20	410	47	0.59
340	35	0.40	369	33	0.39	415	29	0.33	414	43	0.54
344	27	0.30	373	26	0.31	419	23	0.26	417	29	0.36
347	39	0.44	376	38	0.45	422	40	0.46	421	27	0.34
351	44	0.50	380	36	0.42	425	45	0.52	424	41	0.51
354	39	0.44	383	52	0.61	429	40	0.46	427	49	0.61
358	53	0.60	386	64	0.75	432	48	0.55	430	48	0.60
361	59	0.67	389	57	0.67	435	56	0.64	433	54	0.68
364	47	0.53	392	44	0.52	438	60	0.69	436	55	0.69
367	68	0.77	395	70	0.82	440	54	0.62	439	64	0.80
370	62	0.70	398	92	1.08	443	73	0.84	442	67	0.84
373	81	0.91	400	76	0.89	446	56	0.64	444	62	0.78
375	87	0.98	403	70	0.82	449	63	0.72	447	58	0.73
378	89	1.01	406	66	0.77	451	65	0.75	450	68	0.85
381	104	1.17	409	100	1.17	454	79	0.91	452	69	0.86
384	122	1.38	411	98	1.15	457	92	1.06	455	76	0.95
386	111	1.25	414	79	0.93	459	85	0.98	458	105	1.31
389	130	1.47	416	104	1.22	462	87	1.00	460	95	1.19
392	131	1.48	419	104	1.22	465	93	1.07	463	72	0.90
394	100	1.13	422	105	1.23	467	110	1.27	465	91	1.14
397	125	1.41	424	122	1.43	470	122	1.40	468	109	1.36
399	167	1.89	427	124	1.46	472	102	1.17	470	100	1.25
402	155	1.75	429	116	1.36	475	118	1.36	473	116	1.45
404	178	2.01	432	134	1.57	477	156	1.79	475	126	1.58
407	150	1.69	434	158	1.85	480	143	1.64	478	109	1.36
409	165	1.86	437	141	1.65	482	137	1.58	480	136	1.70
412	187	2.11	439	164	1.92	485	157	1.81	483	142	1.78
415	199	2.25	442	142	1.67	488	163	1.87	485	158	1.98
417	209	2.36	444	166	1.95	490	183	2.10	488	147	1.84
420	185	2.09	447	156	1.83	493	167	1.92	490	132	1.65
422	197	2.23	450	175	2.05	495	182	2.09	493	179	2.24
425	201	2.27	452	176	2.07	498	203	2.33	496	161	2.01
427	210	2.37	455	195	2.29	501	209	2.40	498	186	2.33
430	211	2.38	458	160	1.88	503	221	2.54	501	197	2.46
432	212	2.39	460	225	2.64	506	237	2.73	503	209	2.61
435	221	2.50	463	198	2.32	509	226	2.60	506	209	2.61
438	239	2.70	466	206	2.42	511	258	2.97	509	204	2.55

Grade 3			Grade 5			Grade 8			Grade 10		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
440	211	2.38	469	182	2.14	514	244	2.81	512	229	2.86
443	237	2.68	472	225	2.64	517	274	3.15	515	215	2.69
446	253	2.86	475	221	2.59	520	265	3.05	518	240	3.00
449	240	2.71	478	278	3.26	523	278	3.20	521	233	2.91
452	240	2.71	481	215	2.52	526	263	3.02	524	237	2.96
454	232	2.62	484	280	3.29	529	302	3.47	527	238	2.98
457	237	2.68	488	270	3.17	532	299	3.44	530	255	3.19
460	261	2.95	491	240	2.82	535	282	3.24	533	247	3.09
463	225	2.54	495	252	2.96	539	260	2.99	537	238	2.98
467	239	2.70	498	232	2.72	542	254	2.92	540	234	2.93
470	211	2.38	502	227	2.66	546	241	2.77	544	239	2.99
473	228	2.58	506	226	2.65	549	241	2.77	547	217	2.71
477	226	2.55	510	221	2.59	553	227	2.61	551	202	2.53
480	198	2.24	514	233	2.73	557	212	2.44	555	184	2.30
484	180	2.03	519	193	2.27	561	170	1.96	559	158	1.98
487	136	1.54	524	175	2.05	566	129	1.48	564	146	1.83
491	129	1.46	529	168	1.97	570	126	1.45	569	130	1.63
496	108	1.22	534	133	1.56	575	90	1.04	573	104	1.30
500	97	1.10	540	78	0.92	580	65	0.75	579	76	0.95
504	50	0.56	546	79	0.93	586	55	0.63	584	66	0.83
509	51	0.58	553	59	0.69	592	52	0.60	591	60	0.75
514	41	0.46	560	50	0.59	598	26	0.30	597	21	0.26
520	33	0.37	569	31	0.36	605	21	0.24	604	23	0.29
526	21	0.24	578	24	0.28	613	10	0.12	613	16	0.20
532	12	0.14	588	17	0.20	622	13	0.15	622	10	0.13
539	7	0.08	601	5	0.06	633	5	0.06	632	3	0.04
547	4	0.05	616	5	0.06	646	6	0.07	645	4	0.05
556	3	0.03	636	2	0.02	664	1	0.01	661	1	0.01
566	1	0.01	668	1	0.01	693	1	0.01	Total	7995	100.00
Total	8853	100.00	Total	8520	100.00	Total	8695	100.00			

Frequency Distributions of Scale Scores in Mathematics

Grade 3			Grade 5			Grade 8			Grade 10		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
289	3	0.03	321	1	0.01	332	1	0.01	409	4	0.05
296	3	0.03	331	3	0.04	372	4	0.05	419	5	0.06
302	1	0.01	340	1	0.01	380	2	0.02	427	10	0.13
312	4	0.05	347	6	0.07	388	13	0.15	434	20	0.25
316	5	0.06	354	16	0.19	394	23	0.27	440	31	0.39
320	5	0.06	360	19	0.22	400	38	0.44	445	62	0.78
324	14	0.16	366	35	0.41	405	56	0.65	450	75	0.95
328	13	0.15	371	36	0.42	410	65	0.75	455	87	1.10
331	29	0.33	376	58	0.68	414	85	0.98	459	125	1.58
334	23	0.26	380	77	0.90	418	89	1.03	463	132	1.66
337	25	0.28	385	98	1.15	422	128	1.48	467	188	2.37
340	27	0.31	389	97	1.14	426	110	1.27	470	181	2.28
343	42	0.47	393	95	1.12	429	118	1.36	474	199	2.51
346	32	0.36	396	127	1.49	433	159	1.84	477	198	2.50
348	45	0.51	400	112	1.31	436	134	1.55	480	210	2.65
351	46	0.52	403	132	1.55	439	157	1.81	483	212	2.67
354	61	0.69	407	117	1.37	442	159	1.84	486	235	2.96
357	65	0.73	410	135	1.58	445	184	2.13	489	255	3.22
359	62	0.70	413	129	1.51	448	185	2.14	492	248	3.13
362	62	0.70	416	142	1.67	451	169	1.95	495	229	2.89
365	68	0.77	419	145	1.70	454	198	2.29	497	242	3.05
367	104	1.18	422	133	1.56	456	164	1.89	500	250	3.15
370	82	0.93	424	137	1.61	459	176	2.03	503	244	3.08
373	92	1.04	427	151	1.77	461	197	2.28	505	253	3.19
375	88	0.99	430	155	1.82	464	214	2.47	508	226	2.85
378	105	1.19	432	147	1.73	466	228	2.63	510	229	2.89
381	110	1.24	435	160	1.88	469	195	2.25	513	248	3.13
383	135	1.53	437	153	1.80	471	243	2.81	515	224	2.83
386	122	1.38	440	149	1.75	473	221	2.55	518	214	2.70
389	140	1.58	442	168	1.97	476	199	2.30	520	187	2.36
391	161	1.82	445	178	2.09	478	194	2.24	523	183	2.31
394	157	1.77	447	167	1.96	480	192	2.22	525	167	2.11
397	165	1.86	449	203	2.38	482	187	2.16	527	150	1.89
399	164	1.85	452	176	2.07	484	165	1.91	530	157	1.98
402	182	2.06	454	186	2.18	487	205	2.37	532	154	1.94
404	164	1.85	456	181	2.12	489	186	2.15	535	131	1.65
407	191	2.16	459	189	2.22	491	187	2.16	537	124	1.56
410	205	2.32	461	173	2.03	493	207	2.39	539	119	1.50
413	211	2.38	463	194	2.28	496	174	2.01	542	120	1.51
415	218	2.46	465	197	2.31	498	188	2.17	544	116	1.46
418	202	2.28	468	196	2.30	500	166	1.92	546	101	1.27
421	218	2.46	470	190	2.23	502	155	1.79	549	87	1.10
423	227	2.56	472	190	2.23	505	146	1.69	551	89	1.12
426	249	2.81	474	194	2.28	507	120	1.39	554	83	1.05
429	242	2.73	477	176	2.07	509	153	1.77	556	76	0.96
432	209	2.36	479	177	2.08	512	116	1.34	558	88	1.11
435	233	2.63	481	172	2.02	514	146	1.69	561	73	0.92
438	251	2.84	484	163	1.91	516	134	1.55	563	67	0.84

Grade 3			Grade 5			Grade 8			Grade 10		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
440	238	2.69	486	173	2.03	519	111	1.28	566	65	0.82
444	262	2.96	489	140	1.64	521	126	1.46	568	64	0.81
447	242	2.73	491	177	2.08	524	109	1.26	571	44	0.55
450	249	2.81	494	144	1.69	527	119	1.37	573	58	0.73
453	236	2.67	496	149	1.75	529	123	1.42	576	60	0.76
457	231	2.61	499	147	1.73	532	119	1.37	578	48	0.61
460	221	2.50	502	130	1.53	535	103	1.19	581	44	0.55
464	213	2.41	505	146	1.71	538	86	0.99	584	38	0.48
468	238	2.69	508	129	1.51	540	80	0.92	587	45	0.57
473	210	2.37	511	118	1.39	543	97	1.12	589	35	0.44
477	202	2.28	514	120	1.41	547	83	0.96	592	39	0.49
482	177	2.00	517	98	1.15	550	80	0.92	595	35	0.44
488	192	2.17	521	101	1.19	553	66	0.76	598	25	0.32
493	152	1.72	524	99	1.16	557	72	0.83	601	36	0.45
500	157	1.77	528	73	0.86	560	52	0.60	605	28	0.35
507	104	1.18	532	69	0.81	564	45	0.52	608	18	0.23
516	104	1.18	537	68	0.80	568	43	0.50	612	25	0.32
526	48	0.54	542	56	0.66	573	42	0.49	616	28	0.35
539	49	0.55	548	34	0.40	578	43	0.50	620	25	0.32
556	37	0.42	554	31	0.36	584	33	0.38	625	15	0.19
585	22	0.25	562	23	0.27	590	22	0.25	631	12	0.15
614	5	0.06	571	26	0.31	597	19	0.22	638	9	0.11
Total	8851	100.00	583	13	0.15	606	19	0.22	647	6	0.08
			599	11	0.13	618	14	0.16	658	4	0.05
			628	5	0.06	634	10	0.12	673	4	0.05
			656	2	0.02	662	3	0.03	701	9	0.11
			Total	8518	100.00	689	3	0.03	728	2	0.03
						Total	8657	100.00	Total	7929	100.00

Frequency Distributions of Writing Raw Scores

Grade 3			Grade 5			Grade 8			Grade 10		
Raw Score	N.	%	Raw Score	N.	%	Raw Score	N.	%	Raw Score	N.	%
1	28	0.32	1	35	0.41	1	23	0.26	1	41	0.51
2	238	2.70	2	58	0.68	2	60	0.69	2	145	1.80
3	535	6.07	3	268	3.14	3	89	1.02	3	167	2.07
4	850	9.65	4	325	3.81	4	190	2.18	4	308	3.82
5	1745	19.81	5	823	9.64	5	391	4.48	5	491	6.09
6	2554	28.99	6	1275	14.93	6	987	11.31	6	1008	12.50
7	1416	16.07	7	1467	17.18	7	1211	13.87	7	1367	16.96
8	954	10.83	8	1917	22.45	8	2396	27.45	8	2698	33.47
9	376	4.27	9	1465	17.16	9	2284	26.17	9	1161	14.40
10	77	0.87	10	559	6.55	10	761	8.72	10	499	6.19
11	24	0.27	11	254	2.97	11	249	2.85	11	156	1.94
12	10	0.11	12	73	0.85	12	82	0.94	12	17	0.21
13	0	0.00	13	19	0.22	13	5	0.06	13	4	0.05
14	2	0.02	14	1	0.01	14	1	0.01	14	0	0.00
15	1	0.01	15	0	0.00	15	0	0.00	15	0	0.00
Total	8810	100.00	Total	8539	100.00	Total	8729	100.00	Total	8062	100.00

Frequency Distributions of Sub-Writing Raw Scores

Stand-alone Raw Score	Grade 3 N.	%	Stand-alone Raw Score	Grade 5 N.	%	Stand-alone Raw Score	Grade 8 N.	%	Stand-alone Raw Score	Grade 10 N.	%
0	80	0.90	0	44	0.52	0	15	0.17	0	47	0.59
2	669	7.56	2	345	4.05	2	111	1.28	2	240	3.00
3	897	10.13	3	356	4.18	3	174	2.00	3	281	3.52
4	4204	47.50	4	1904	22.34	4	1330	15.31	4	1172	14.67
5	1693	19.13	5	1653	19.40	5	1562	17.98	5	1499	18.77
6	1115	12.60	6	3069	36.01	6	4553	52.41	6	3480	43.57
7	156	1.76	7	768	9.01	7	736	8.47	7	905	11.33
8	30	0.34	8	327	3.84	8	198	2.28	8	352	4.41
9	4	0.05	9	50	0.59	9	8	0.09	9	8	0.10
10	3	0.03	10	6	0.07	10	1	0.01	10	3	0.04
Total	8851	100.00	Total	8522	100.00	Total	8688	100.00	Total	7987	100.00

Text-based Raw Score	Grade 3 N.	%	Text-based Raw Score	Grade 5 N.	%	Text-based Raw Score	Grade 8 N.	%	Text-based Raw Score	Grade 10 N.	%
0	697	8.03	0	25	0.30	0	36	0.42	0	59	0.77
1	2600	29.95	1	1711	20.35	1	719	8.45	1	1092	14.18
2	4507	51.92	2	4120	49.01	2	3977	46.76	2	5287	68.64
3	836	9.63	3	2378	28.29	3	3120	36.68	3	1237	16.06
4	37	0.43	4	170	2.02	4	653	7.68	4	25	0.32
5	3	0.03	5	2	0.02	5	1	0.01	5	3	0.04
Total	8680	100.00	Total	8406	100.00	Total	8506	100.00	Total	7703	100.00

Frequency Distributions of Scale Scores in Science

Grade 4			Grade 6			Grade 8			Grade 11		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
231	1	0.01	198	1	0.011	162	3	0.035	191	4	0.064
249	1	0.01	212	1	0.011	181	5	0.058	207	9	0.145
253	2	0.02	226	2	0.022	202	11	0.128	223	6	0.097
256	4	0.04	235	2	0.022	214	8	0.093	232	15	0.242
259	8	0.09	241	1	0.011	222	12	0.14	239	17	0.274
261	10	0.11	246	2	0.022	229	18	0.21	245	35	0.564
264	5	0.06	251	4	0.045	235	47	0.549	250	41	0.661
266	7	0.08	255	4	0.045	240	56	0.654	254	67	1.08
268	12	0.13	258	17	0.189	245	71	0.829	257	87	1.403
270	15	0.17	261	19	0.212	249	126	1.472	261	81	1.306
272	13	0.15	264	26	0.29	252	117	1.367	264	104	1.677
274	23	0.26	266	39	0.434	256	154	1.799	266	128	2.064
276	12	0.13	269	43	0.479	259	174	2.032	269	122	1.967
277	27	0.30	271	54	0.601	262	173	2.021	272	155	2.499
279	42	0.47	273	59	0.657	265	178	2.079	274	174	2.806
280	26	0.29	275	78	0.869	268	224	2.617	276	157	2.531
282	32	0.36	277	77	0.858	270	223	2.605	278	162	2.612
283	39	0.44	279	105	1.17	273	217	2.535	280	174	2.806
285	40	0.45	281	106	1.181	275	224	2.617	282	184	2.967
286	59	0.66	283	127	1.415	277	242	2.827	284	194	3.128
288	54	0.61	285	143	1.593	280	268	3.13	286	168	2.709
289	86	0.97	286	138	1.537	282	265	3.095	288	215	3.467
291	81	0.91	288	134	1.493	284	236	2.757	290	218	3.515
292	80	0.90	289	187	2.083	286	250	2.92	292	164	2.644
293	76	0.85	291	199	2.217	288	266	3.107	293	181	2.918
295	87	0.98	293	210	2.339	290	253	2.955	295	190	3.064
296	103	1.16	294	195	2.172	292	278	3.247	297	195	3.144
297	124	1.39	296	221	2.462	294	252	2.944	298	160	2.58
299	160	1.80	297	259	2.885	296	286	3.341	300	186	2.999
300	167	1.87	299	263	2.929	298	282	3.294	302	159	2.564
301	180	2.02	300	255	2.84	300	249	2.909	303	181	2.918
303	182	2.04	301	256	2.851	302	248	2.897	305	154	2.483
304	210	2.36	303	311	3.464	304	243	2.838	306	151	2.435
305	226	2.54	304	288	3.208	306	247	2.885	308	173	2.789
307	229	2.57	306	307	3.419	308	260	3.037	309	139	2.241
308	278	3.12	307	302	3.364	309	261	3.049	311	155	2.499
309	286	3.21	309	304	3.386	311	192	2.243	313	153	2.467
311	328	3.68	310	324	3.609	313	184	2.149	314	135	2.177
312	319	3.58	311	307	3.419	315	193	2.254	316	119	1.919
313	328	3.68	313	328	3.653	317	186	2.173	317	142	2.29
315	346	3.88	314	345	3.843	319	175	2.044	319	115	1.854
316	362	4.06	316	305	3.397	321	150	1.752	320	107	1.725
317	345	3.87	317	295	3.286	323	147	1.717	322	100	1.612
319	339	3.80	319	285	3.174	325	129	1.507	323	91	1.467
320	342	3.84	320	279	3.108	327	119	1.39	325	80	1.29
322	356	4.00	322	262	2.918	329	109	1.273	327	57	0.919
323	370	4.15	323	233	2.595	331	97	1.133	328	61	0.984
325	377	4.23	325	234	2.606	333	87	1.016	330	57	0.919

Grade 4			Grade 6			Grade 8			Grade 11		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
327	324	3.64	327	205	2.283	336	73	0.853	332	35	0.564
328	279	3.13	328	179	1.994	338	64	0.748	333	43	0.693
330	296	3.32	330	156	1.738	340	41	0.479	335	39	0.629
332	237	2.66	332	121	1.348	343	46	0.537	337	41	0.661
334	225	2.53	334	91	1.014	345	44	0.514	339	25	0.403
336	202	2.27	335	85	0.947	348	23	0.269	341	24	0.387
338	168	1.89	337	58	0.646	350	18	0.21	343	17	0.274
341	108	1.21	339	42	0.468	353	24	0.28	345	12	0.193
343	94	1.05	342	40	0.446	356	9	0.105	348	19	0.306
346	65	0.73	344	24	0.267	359	9	0.105	350	5	0.081
349	40	0.45	346	19	0.212	363	8	0.093	353	4	0.064
353	42	0.47	349	9	0.1	367	5	0.058	356	5	0.081
358	19	0.21	352	3	0.033	371	1	0.012	359	3	0.048
363	11	0.12	355	6	0.067	380	1	0.012	362	2	0.032
371	1	0.01	359	2	0.022	Total	8561	100	366	2	0.032
Total	8910	100.00	363	1	0.011				371	2	0.032
			368	1	0.011				393	2	0.032
			Total	8978	100				Total	6202	100

Frequency Distributions of Scale Scores in Science

Grade 4			Grade 6			Grade 8			Grade 11		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
231	1	0.01	198	1	0.011	162	3	0.035	191	4	0.064
249	1	0.01	212	1	0.011	181	5	0.058	207	9	0.145
253	2	0.02	226	2	0.022	202	11	0.128	223	6	0.097
256	4	0.04	235	2	0.022	214	8	0.093	232	15	0.242
259	8	0.09	241	1	0.011	222	12	0.14	239	17	0.274
261	10	0.11	246	2	0.022	229	18	0.21	245	35	0.564
264	5	0.06	251	4	0.045	235	47	0.549	250	41	0.661
266	7	0.08	255	4	0.045	240	56	0.654	254	67	1.08
268	12	0.13	258	17	0.189	245	71	0.829	257	87	1.403
270	15	0.17	261	19	0.212	249	126	1.472	261	81	1.306
272	13	0.15	264	26	0.29	252	117	1.367	264	104	1.677
274	23	0.26	266	39	0.434	256	154	1.799	266	128	2.064
276	12	0.13	269	43	0.479	259	174	2.032	269	122	1.967
277	27	0.30	271	54	0.601	262	173	2.021	272	155	2.499
279	42	0.47	273	59	0.657	265	178	2.079	274	174	2.806
280	26	0.29	275	78	0.869	268	224	2.617	276	157	2.531
282	32	0.36	277	77	0.858	270	223	2.605	278	162	2.612
283	39	0.44	279	105	1.17	273	217	2.535	280	174	2.806
285	40	0.45	281	106	1.181	275	224	2.617	282	184	2.967
286	59	0.66	283	127	1.415	277	242	2.827	284	194	3.128
288	54	0.61	285	143	1.593	280	268	3.13	286	168	2.709
289	86	0.97	286	138	1.537	282	265	3.095	288	215	3.467
291	81	0.91	288	134	1.493	284	236	2.757	290	218	3.515
292	80	0.90	289	187	2.083	286	250	2.92	292	164	2.644
293	76	0.85	291	199	2.217	288	266	3.107	293	181	2.918
295	87	0.98	293	210	2.339	290	253	2.955	295	190	3.064
296	103	1.16	294	195	2.172	292	278	3.247	297	195	3.144
297	124	1.39	296	221	2.462	294	252	2.944	298	160	2.58
299	160	1.80	297	259	2.885	296	286	3.341	300	186	2.999
300	167	1.87	299	263	2.929	298	282	3.294	302	159	2.564
301	180	2.02	300	255	2.84	300	249	2.909	303	181	2.918
303	182	2.04	301	256	2.851	302	248	2.897	305	154	2.483
304	210	2.36	303	311	3.464	304	243	2.838	306	151	2.435
305	226	2.54	304	288	3.208	306	247	2.885	308	173	2.789
307	229	2.57	306	307	3.419	308	260	3.037	309	139	2.241
308	278	3.12	307	302	3.364	309	261	3.049	311	155	2.499
309	286	3.21	309	304	3.386	311	192	2.243	313	153	2.467
311	328	3.68	310	324	3.609	313	184	2.149	314	135	2.177
312	319	3.58	311	307	3.419	315	193	2.254	316	119	1.919
313	328	3.68	313	328	3.653	317	186	2.173	317	142	2.29
315	346	3.88	314	345	3.843	319	175	2.044	319	115	1.854
316	362	4.06	316	305	3.397	321	150	1.752	320	107	1.725
317	345	3.87	317	295	3.286	323	147	1.717	322	100	1.612
319	339	3.80	319	285	3.174	325	129	1.507	323	91	1.467
320	342	3.84	320	279	3.108	327	119	1.39	325	80	1.29
322	356	4.00	322	262	2.918	329	109	1.273	327	57	0.919
323	370	4.15	323	233	2.595	331	97	1.133	328	61	0.984
325	377	4.23	325	234	2.606	333	87	1.016	330	57	0.919

Grade 4			Grade 6			Grade 8			Grade 11		
Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%	Scale Score	N.	%
327	324	3.64	327	205	2.283	336	73	0.853	332	35	0.564
328	279	3.13	328	179	1.994	338	64	0.748	333	43	0.693
330	296	3.32	330	156	1.738	340	41	0.479	335	39	0.629
332	237	2.66	332	121	1.348	343	46	0.537	337	41	0.661
334	225	2.53	334	91	1.014	345	44	0.514	339	25	0.403
336	202	2.27	335	85	0.947	348	23	0.269	341	24	0.387
338	168	1.89	337	58	0.646	350	18	0.21	343	17	0.274
341	108	1.21	339	42	0.468	353	24	0.28	345	12	0.193
343	94	1.05	342	40	0.446	356	9	0.105	348	19	0.306
346	65	0.73	344	24	0.267	359	9	0.105	350	5	0.081
349	40	0.45	346	19	0.212	363	8	0.093	353	4	0.064
353	42	0.47	349	9	0.1	367	5	0.058	356	5	0.081
358	19	0.21	352	3	0.033	371	1	0.012	359	3	0.048
363	11	0.12	355	6	0.067	380	1	0.012	362	2	0.032
371	1	0.01	359	2	0.022	Total	8561	100	366	2	0.032
Total	8910	100.00	363	1	0.011				371	2	0.032
			368	1	0.011				393	2	0.032
			Total	8978	100				Total	6202	100

Appendix I

Conversion Tables from Raw Scores to Scale Scores

2001 DSTP RAW SCORE TO SCALED SCORE CONVERSION TABLES

	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS
GR 3 READING	000-164	001-194	002-227	003-247	004-261	005-273	006-283	007-291	008-299	009-306
	010-312	011-317	012-322	013-327	014-332	015-336	016-340	017-344	018-347	019-351
	020-354	021-358	022-361	023-364	024-367	025-370	026-373	027-375	028-378	029-381
	030-384	031-386	032-389	033-392	034-394	035-397	036-399	037-402	038-404	039-407
	040-409	041-412	042-415	043-417	044-420	045-422	046-425	047-427	048-430	049-432
	050-435	051-438	052-440	053-443	054-446	055-449	056-452	057-454	058-457	059-460
	060-463	061-467	062-470	063-473	064-477	065-480	066-484	067-487	068-491	069-496
	070-500	071-504	072-509	073-514	074-520	075-526	076-532	077-539	078-547	079-556
	080-566	081-580	082-597	083-627	084-655					
GR 3 MATH	000-174	001-203	002-232	003-250	004-263	005-273	006-282	007-289	008-296	009-302
	010-307	011-312	012-316	013-320	014-324	015-328	016-331	017-334	018-337	019-340
	020-343	021-346	022-348	023-351	024-354	025-357	026-359	027-362	028-365	029-367
	030-370	031-373	032-375	033-378	034-381	035-383	036-386	037-389	038-391	039-394
	040-397	041-399	042-402	043-404	044-407	045-410	046-413	047-415	048-418	049-421
	050-423	051-426	052-429	053-432	054-435	055-438	056-440	057-444	058-447	059-450
	060-453	061-457	062-460	063-464	064-468	065-473	066-477	067-482	068-488	069-493
	070-500	071-507	072-516	073-526	074-539	075-556	076-585	077-614		
GR 5 READING	000-213	001-241	002-270	003-288	004-300	005-310	006-319	007-326	008-333	009-338
	010-344	011-349	012-353	013-358	014-362	015-366	016-369	017-373	018-376	019-380
	020-383	021-386	022-389	023-392	024-395	025-398	026-400	027-403	028-406	029-409
	030-411	031-414	032-416	033-419	034-422	035-424	036-427	037-429	038-432	039-434
	040-437	041-439	042-442	043-444	044-447	045-450	046-452	047-455	048-458	049-460
	050-463	051-466	052-469	053-472	054-475	055-478	056-481	057-484	058-488	059-491
	060-495	061-498	062-502	063-506	064-510	065-514	066-519	067-524	068-529	069-534
	070-540	071-546	072-553	073-560	074-569	075-578	076-588	077-601	078-616	079-636
	080-668	081-698								
GR 5 MATH	000-233	001-261	002-290	003-308	004-321	005-331	006-340	007-347	008-354	009-360
	010-366	011-371	012-376	013-380	014-385	015-389	016-393	017-396	018-400	019-403
	020-407	021-410	022-413	023-416	024-419	025-422	026-424	027-427	028-430	029-432
	030-435	031-437	032-440	033-442	034-445	035-447	036-449	037-452	038-454	039-456
	040-459	041-461	042-463	043-465	044-468	045-470	046-472	047-474	048-477	049-479
	050-481	051-484	052-486	053-489	054-491	055-494	056-496	057-499	058-502	059-505
	060-508	061-511	062-514	063-517	064-521	065-524	066-528	067-532	068-537	069-542
	070-548	071-554	072-562	073-571	074-583	075-599	076-628	077-656		

2001 DSTP RAW SCORE TO SCALED SCORE CONVERSION TABLES

	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS
GR 8 READING	000-256	001-256	002-256	003-285	004-313	005-331	006-343	007-353	008-362	009-369
	010-375	011-381	012-386	013-391	014-396	015-400	016-404	017-408	018-412	019-415
	020-419	021-422	022-425	023-429	024-432	025-435	026-438	027-440	028-443	029-446
	030-449	031-451	032-454	033-457	034-459	035-462	036-465	037-467	038-470	039-472
	040-475	041-477	042-480	043-482	044-485	045-488	046-490	047-493	048-495	049-498
	050-501	051-503	052-506	053-509	054-511	055-514	056-517	057-520	058-523	059-526
	060-529	061-532	062-535	063-539	064-542	065-546	066-549	067-553	068-557	069-561
	070-566	071-570	072-575	073-580	074-586	075-592	076-598	077-605	078-613	079-622
	080-633	081-646	082-664	083-693	084-721					
GR 8 MATH	000-275	001-304	002-332	003-350	004-362	005-372	006-380	007-388	008-394	009-400
	010-405	011-410	012-414	013-418	014-422	015-426	016-429	017-433	018-436	019-439
	020-442	021-445	022-448	023-451	024-454	025-456	026-459	027-461	028-464	029-466
	030-469	031-471	032-473	033-476	034-478	035-480	036-482	037-484	038-487	039-489
	040-491	041-493	042-496	043-498	044-500	045-502	046-505	047-507	048-509	049-512
	050-514	051-516	052-519	053-521	054-524	055-527	056-529	057-532	058-535	059-538
	060-540	061-543	062-547	063-550	064-553	065-557	066-560	067-564	068-568	069-573
	070-578	071-584	072-590	073-597	074-606	075-618	076-634	077-662	078-689	
GR 10 READING	000-264	001-292	002-321	003-338	004-351	005-360	006-369	007-376	008-382	009-388
	010-393	011-398	012-402	013-406	014-410	015-414	016-417	017-421	018-424	019-427
	020-430	021-433	022-436	023-439	024-442	025-444	026-447	027-450	028-452	029-455
	030-458	031-460	032-463	033-465	034-468	035-470	036-473	037-475	038-478	039-480
	040-483	041-485	042-488	043-490	044-493	045-496	046-498	047-501	048-503	049-506
	050-509	051-512	052-515	053-518	054-521	055-524	056-527	057-530	058-533	059-537
	060-540	061-544	062-547	063-551	064-555	065-559	066-564	067-569	068-573	069-579
	070-584	071-591	072-597	073-604	074-613	075-622	076-632	077-645	078-661	079-681
	080-714	081-745								
GR 10 MATH	000-324	001-352	002-380	003-397	004-409	005-419	006-427	007-434	008-440	009-445
	010-450	011-455	012-459	013-463	014-467	015-470	016-474	017-477	018-480	019-483
	020-486	021-489	022-492	023-495	024-497	025-500	026-503	027-505	028-508	029-510
	030-513	031-515	032-518	033-520	034-523	035-525	036-527	037-530	038-532	039-535
	040-537	041-539	042-542	043-544	044-546	045-549	046-551	047-554	048-556	049-558
	050-561	051-563	052-566	053-568	054-571	055-573	056-576	057-578	058-581	059-584
	060-587	061-589	062-592	063-595	064-598	065-601	066-605	067-608	068-612	069-616
	070-620	071-625	072-631	073-638	074-647	075-658	076-673	077-701	078-728	

2001 DSTP RAW SCORE TO SCALED SCORE CONVERSION TABLES

GR 4 SCIENCE

RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS	RS-SS
000-206	001-219	002-231	003-239	004-244	005-249	006-253	007-256	008-259	009-261
010-264	011-266	012-268	013-270	014-272	015-274	016-276	017-277	018-279	019-280
020-282	021-283	022-285	023-286	024-288	025-289	026-291	027-292	028-293	029-295
030-296	031-297	032-299	033-300	034-301	035-303	036-304	037-305	038-307	039-308
040-309	041-311	042-312	043-313	044-315	045-316	046-317	047-319	048-320	049-322
050-323	051-325	052-327	053-328	054-330	055-332	056-334	057-336	058-338	059-341
060-343	061-346	062-349	063-353	064-358	065-363	066-371	067-384	068-397	

GR 4 SOCIAL STUDIES

000-190	001-204	002-218	003-226	004-232	005-238	006-242	007-246	008-249	009-252
010-255	011-257	012-260	013-262	014-264	015-267	016-269	017-271	018-273	019-274
020-276	021-278	022-280	023-281	024-283	025-285	026-286	027-288	028-289	029-291
030-292	031-294	032-296	033-297	034-299	035-300	036-301	037-303	038-304	039-306
040-307	041-309	042-310	043-312	044-314	045-315	046-317	047-318	048-320	049-322
050-323	051-325	052-327	053-329	054-331	055-333	056-335	057-337	058-339	059-342
060-345	061-348	062-351	063-355	064-360	065-366	066-374	067-387	068-401	

GR 6 SCIENCE

000-198	001-212	002-226	003-235	004-241	005-246	006-251	007-255	008-258	009-261
010-264	011-266	012-269	013-271	014-273	015-275	016-277	017-279	018-281	019-283
020-285	021-286	022-288	023-289	024-291	025-293	026-294	027-296	028-297	029-299
030-300	031-301	032-303	033-304	034-306	035-307	036-309	037-310	038-311	039-313
040-314	041-316	042-317	043-319	044-320	045-322	046-323	047-325	048-327	049-328
050-330	051-332	052-334	053-335	054-337	055-339	056-342	057-344	058-346	059-349
060-352	061-355	062-359	063-363	064-368	065-374	066-383	067-397	068-411	

GR 6 SOCIAL STUDIES

000-208	001-220	002-233	003-241	004-246	005-251	006-255	007-258	008-261	009-264
010-266	011-268	012-270	013-272	014-274	015-276	016-278	017-280	018-281	019-283
020-284	021-286	022-287	023-289	024-290	025-292	026-293	027-295	028-296	029-297
030-299	031-300	032-301	033-303	034-304	035-305	036-307	037-308	038-309	039-311
040-312	041-313	042-315	043-316	044-318	045-319	046-320	047-322	048-323	049-325
050-327	051-328	052-330	053-332	054-333	055-335	056-337	057-340	058-342	059-344
060-347	061-350	062-353	063-357	064-361	065-367	066-374	067-387	068-399	

2001 DSTP RAW SCORE TO SCALED SCORE CONVERSION TABLES

RS-SS RS-SS RS-SS RS-SS RS-SS RS-SS RS-SS RS-SS RS-SS RS-SS

GR 8 SCIENCE

000-162 001-181 002-202 003-214 004-222 005-229 006-235 007-240 008-245 009-249
010-252 011-256 012-259 013-262 014-265 015-268 016-270 017-273 018-275 019-277
020-280 021-282 022-284 023-286 024-288 025-290 026-292 027-294 028-296 029-298
030-300 031-302 032-304 033-306 034-308 035-309 036-311 037-313 038-315 039-317
040-319 041-321 042-323 043-325 044-327 045-329 046-331 047-333 048-336 049-338
050-340 051-343 052-345 053-348 054-350 055-353 056-356 057-359 058-363 059-367
060-371 061-375 062-380 063-386 064-393 065-402 066-414 067-434 068-454

GR 8 SOCIAL STUDIES

000-182 001-198 002-215 003-226 004-234 005-240 006-245 007-249 008-253 009-257
010-260 011-263 012-266 013-269 014-271 015-274 016-276 017-278 018-280 019-282
020-284 021-286 022-288 023-290 024-292 025-293 026-295 027-297 028-298 029-300
030-302 031-303 032-305 033-306 034-308 035-309 036-311 037-312 038-314 039-316
040-317 041-319 042-320 043-322 044-323 045-325 046-327 047-328 048-330 049-332
050-334 051-335 052-337 053-339 054-341 055-344 056-346 057-348 058-351 059-354
060-357 061-360 062-364 063-368 064-374 065-381 066-390 067-406 068-422

GR 11 SCIENCE

000-191 001-207 002-223 003-232 004-239 005-245 006-250 007-254 008-257 009-261
010-264 011-266 012-269 013-272 014-274 015-276 016-278 017-280 018-282 019-284
020-286 021-288 022-290 023-292 024-293 025-295 026-297 027-298 028-300 029-302
030-303 031-305 032-306 033-308 034-309 035-311 036-313 037-314 038-316 039-317
040-319 041-320 042-322 043-323 044-325 045-327 046-328 047-330 048-332 049-333
050-335 051-337 052-339 053-341 054-343 055-345 056-348 057-350 058-353 059-356
060-359 061-362 062-366 063-371 064-377 065-384 066-393 067-410 068-425

GR 11 SOCIAL STUDIES

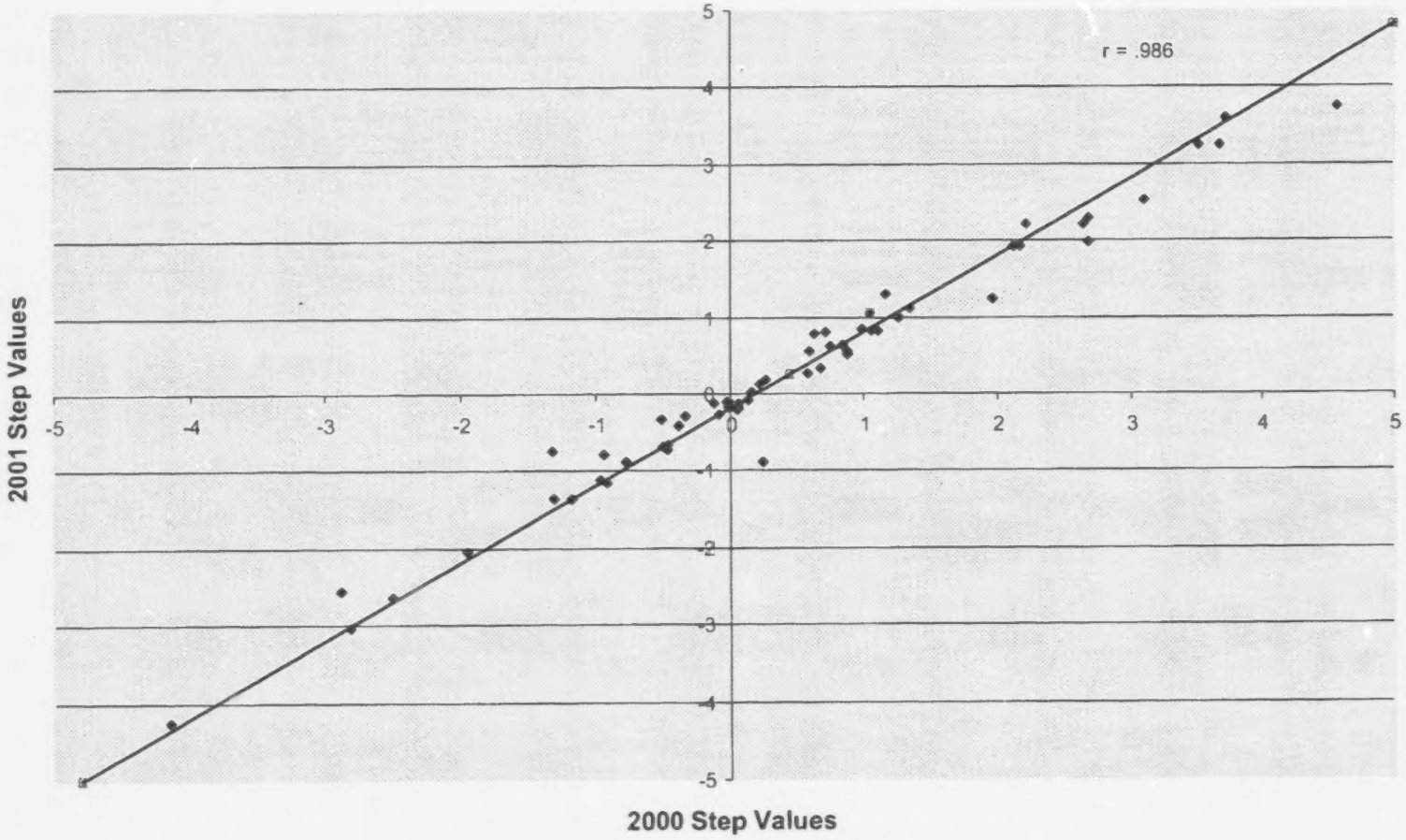
000-170 001-188 002-207 003-218 004-226 005-233 006-238 007-243 008-248 009-252
010-255 011-259 012-262 013-265 014-268 015-271 016-273 017-276 018-278 019-281
020-283 021-285 022-287 023-290 024-292 025-294 026-296 027-298 028-300 029-302
030-304 031-306 032-308 033-310 034-312 035-314 036-315 037-317 038-319 039-321
040-323 041-325 042-327 043-329 044-331 045-333 046-335 047-337 048-339 049-341
050-343 051-345 052-348 053-350 054-352 055-355 056-358 057-361 058-364 059-367
060-371 061-374 062-379 063-384 064-390 065-398 066-409 067-427 068-445

Attachment J

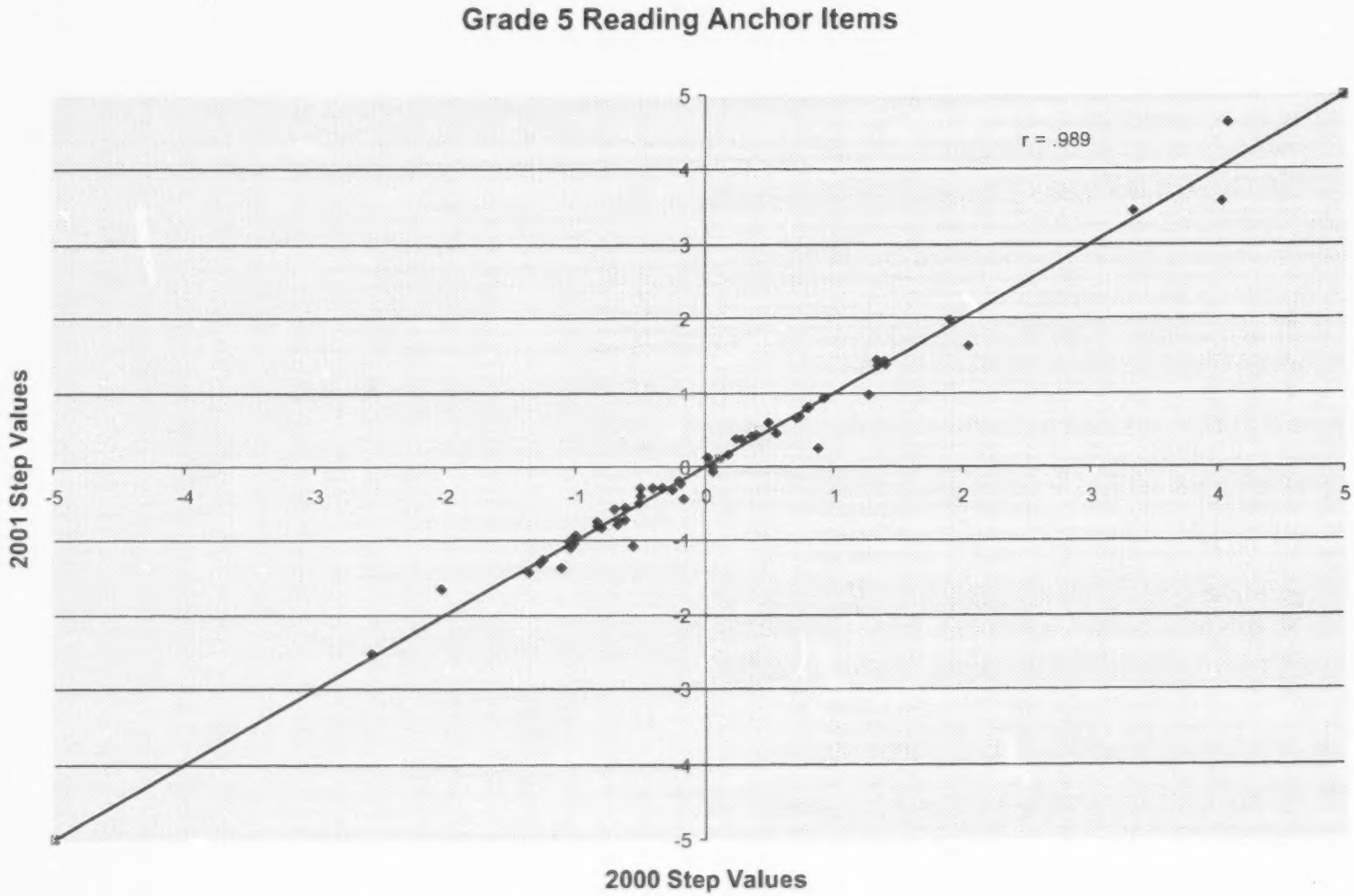
Comparisons of Step-Values for Anchor Items by Test

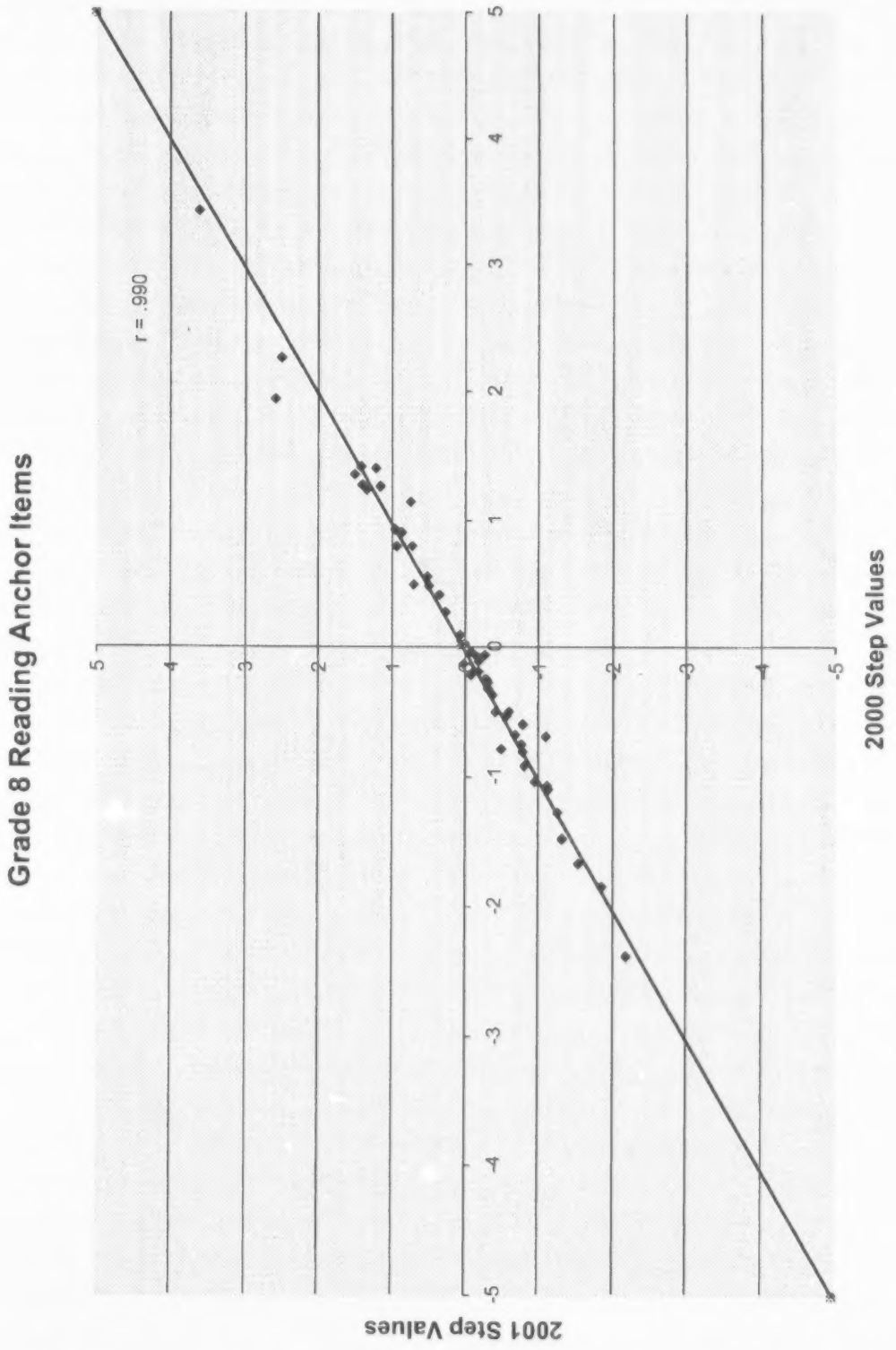
BEST COPY AVAILABLE

Grade 3 Reading Anchor Items

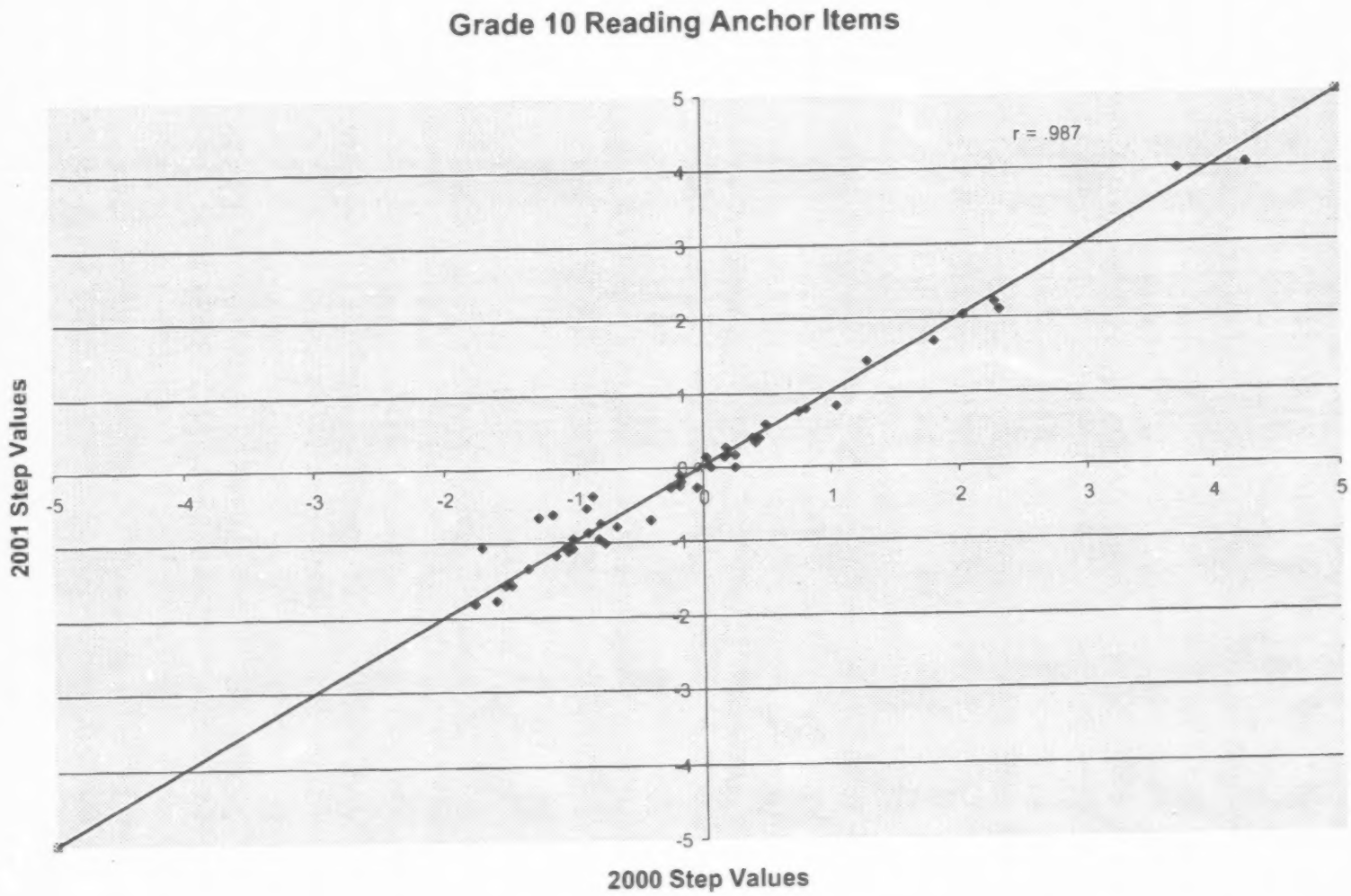


BEST COPY AVAILABLE

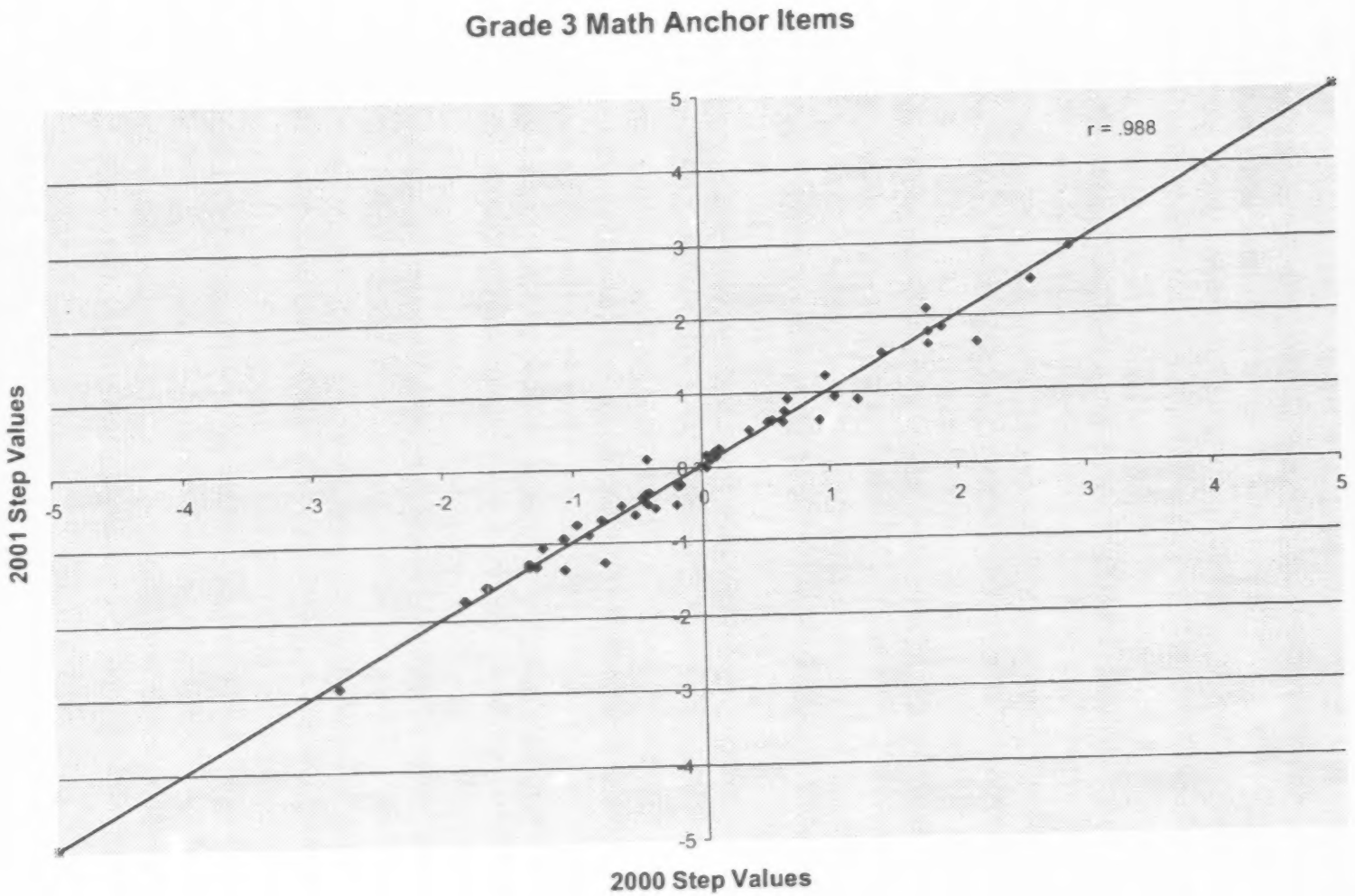




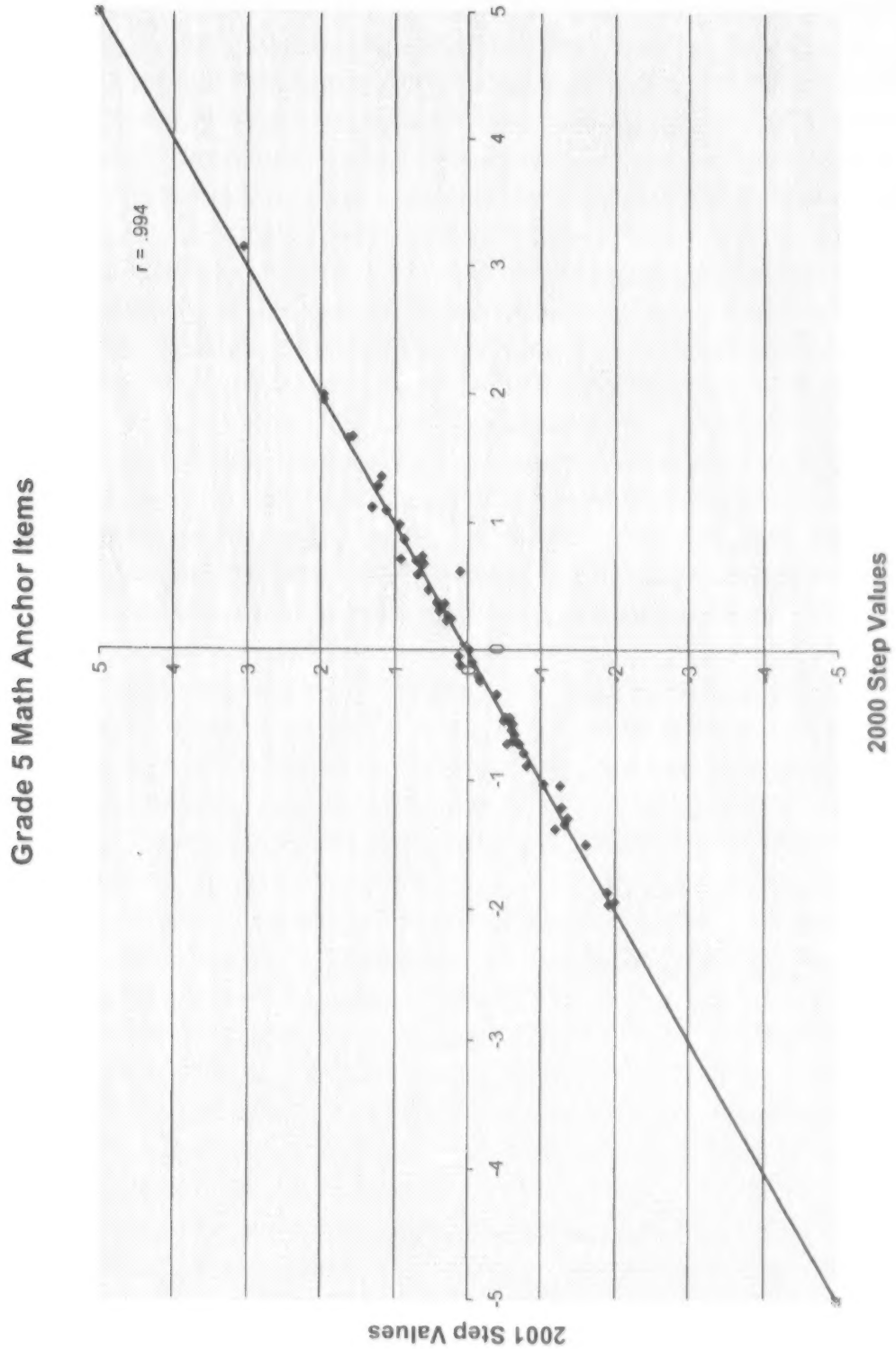
BEST COPY AVAILABLE



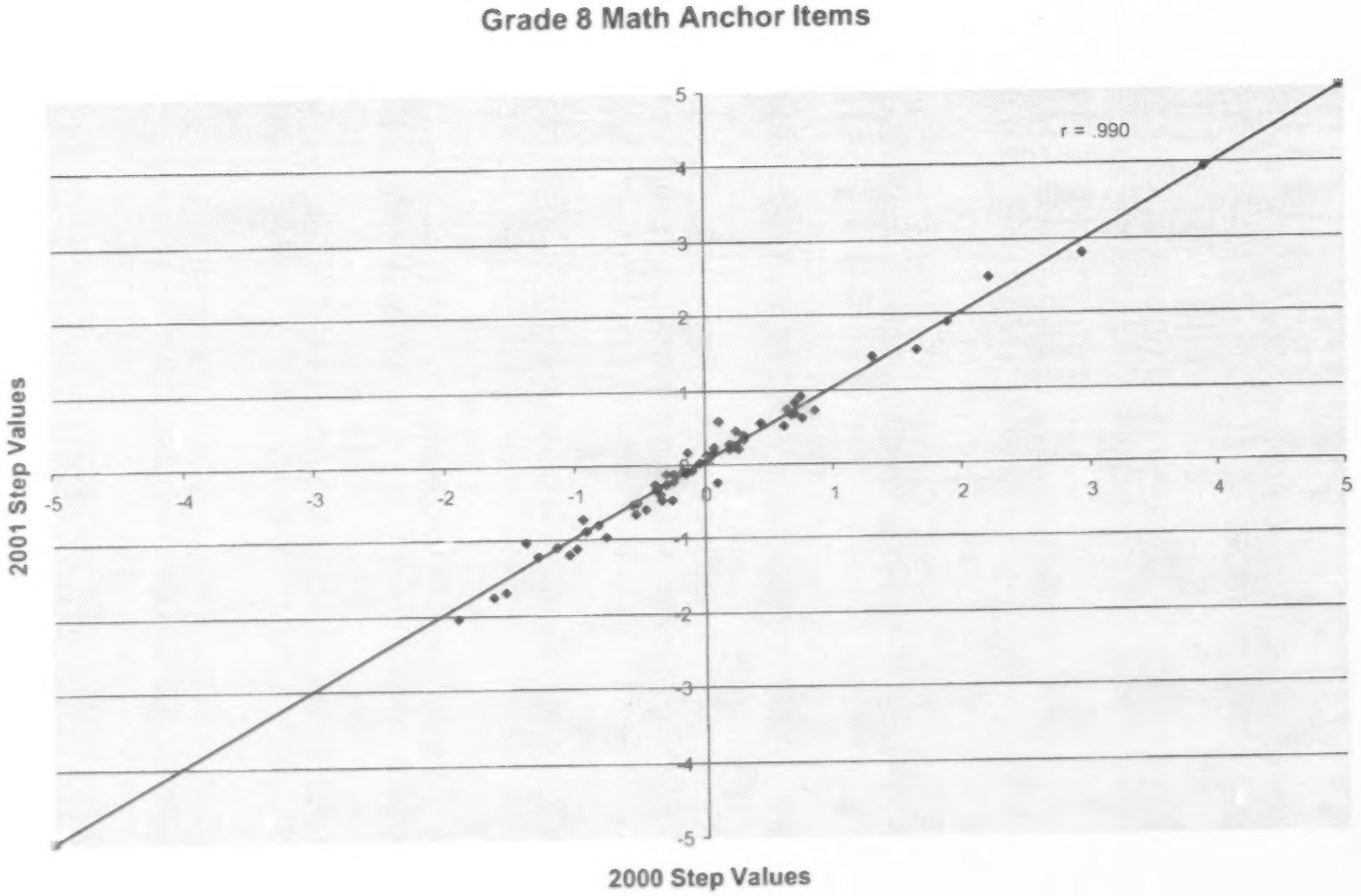
BEST COPY AVAILABLE



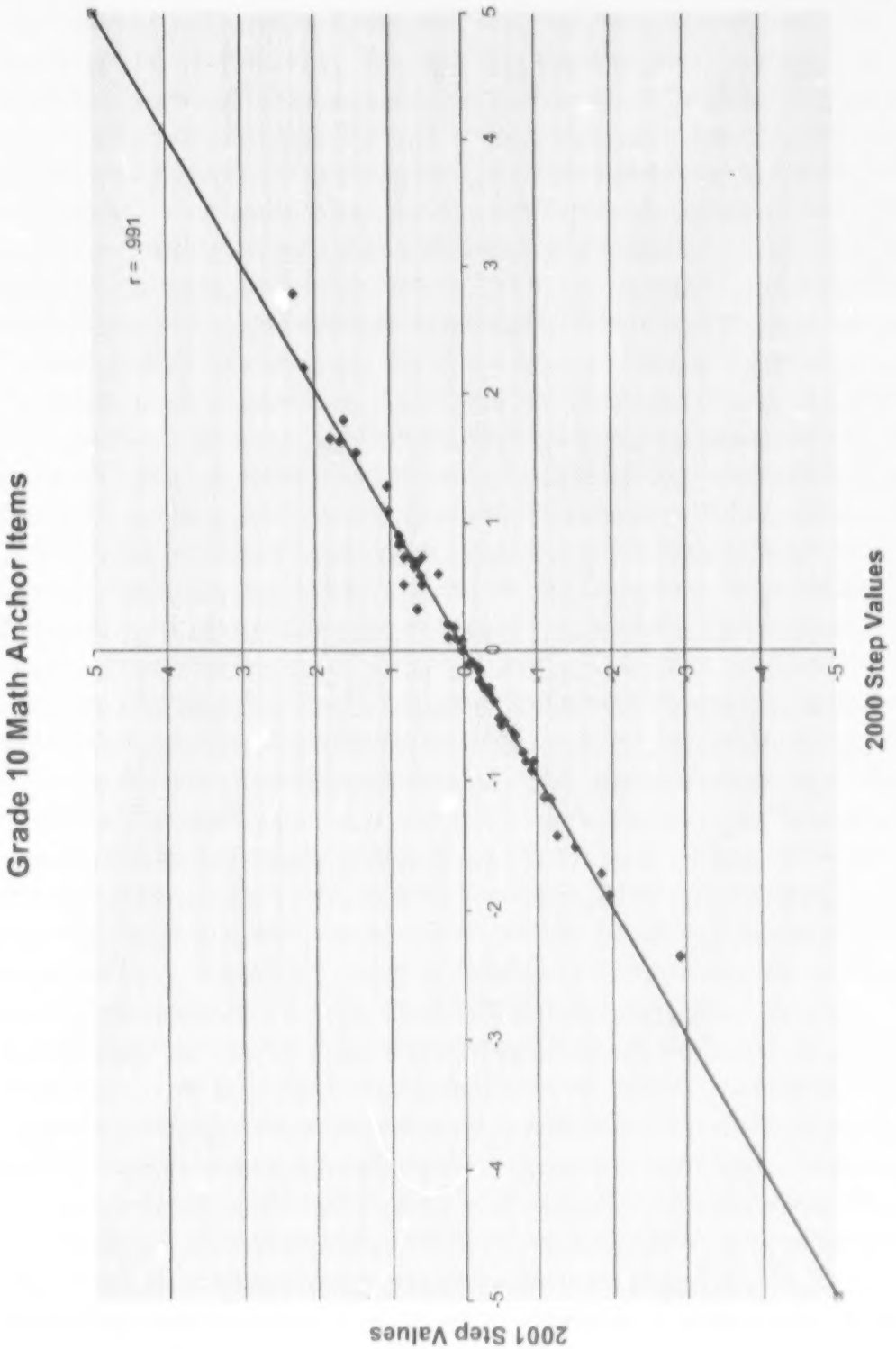
BEST COPY AVAILABLE



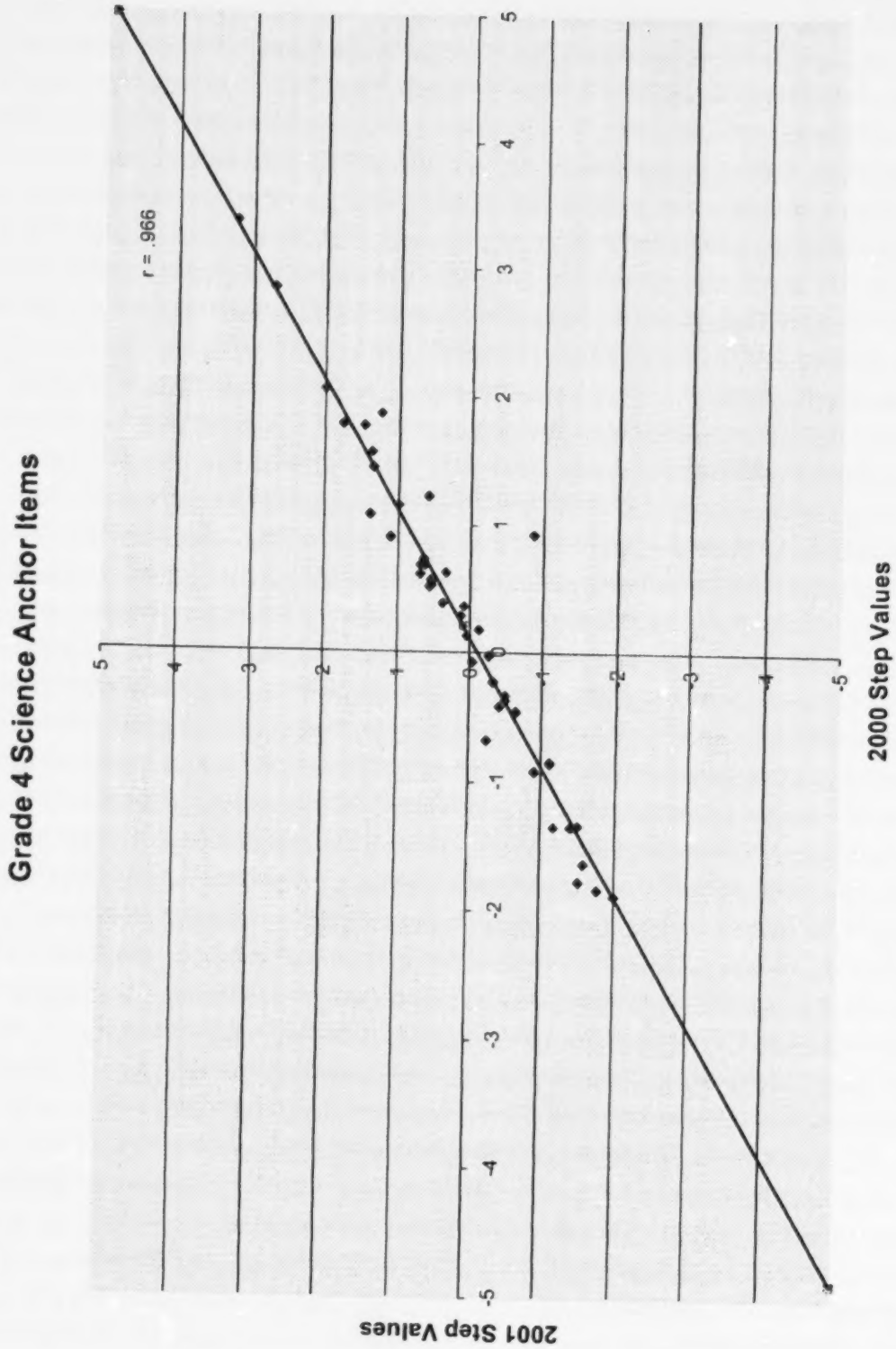
BEST COPY AVAILABLE



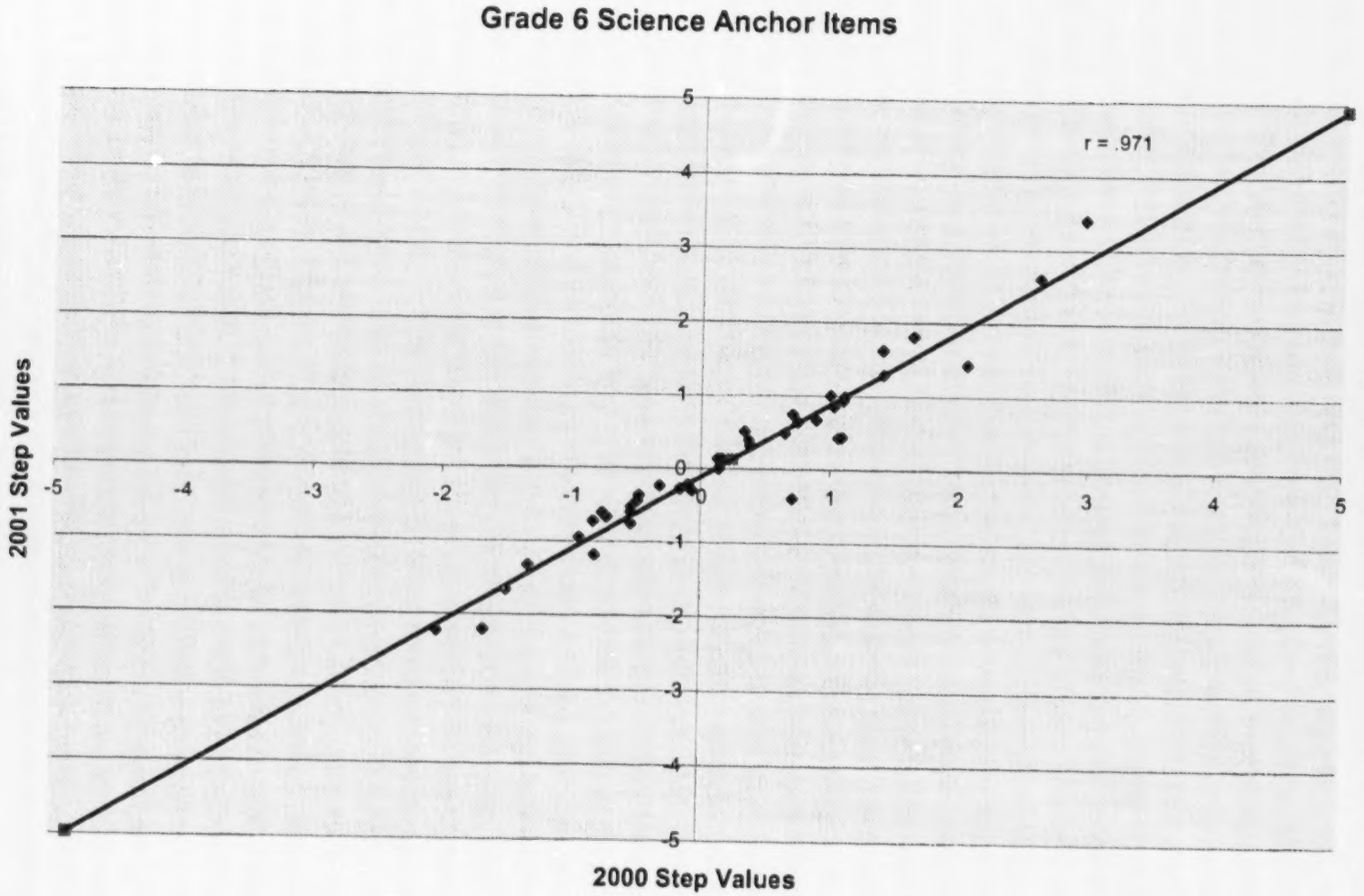
BEST COPY AVAILABLE

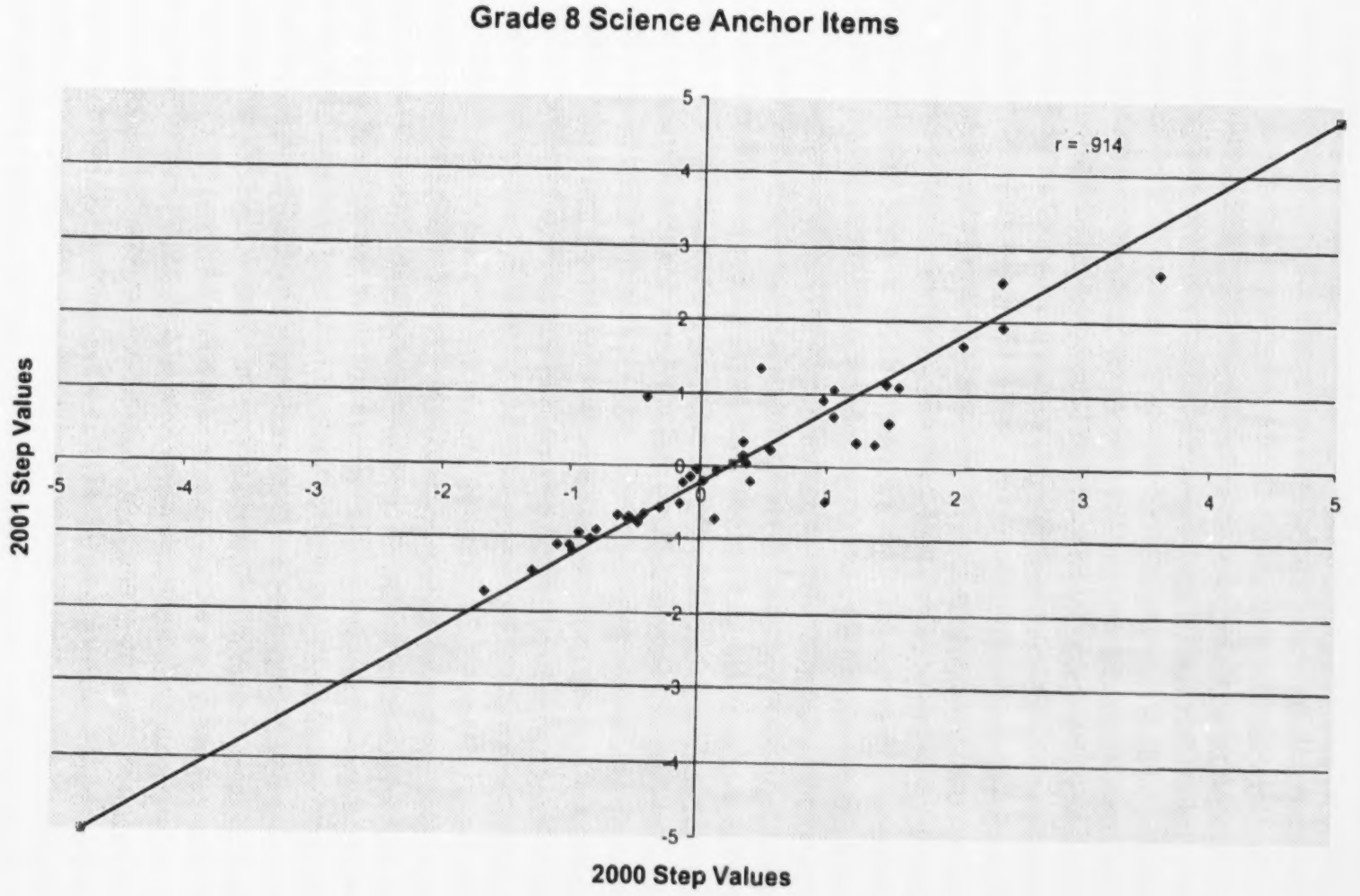


BEST COPY AVAILABLE

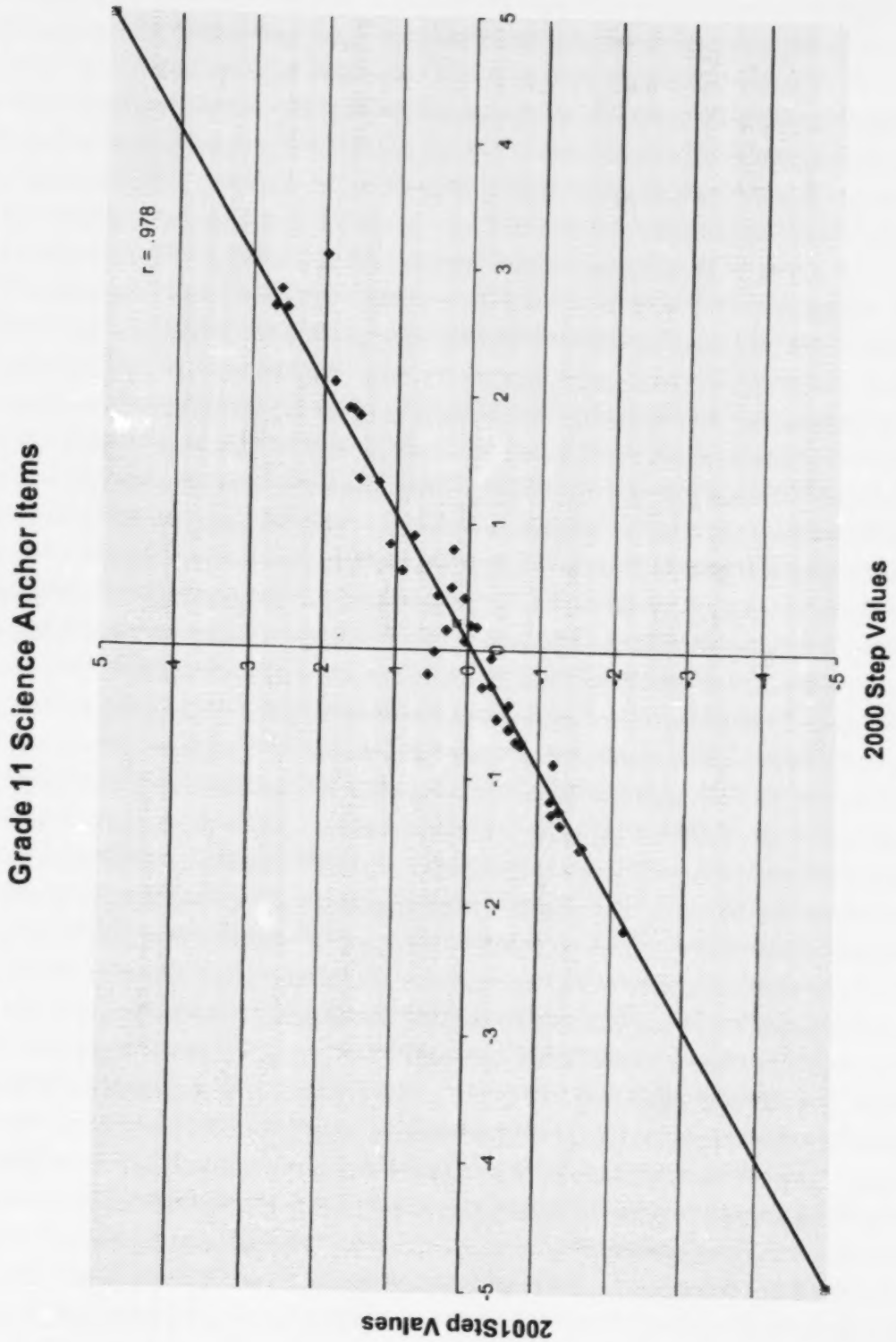


BEST COPY AVAILABLE

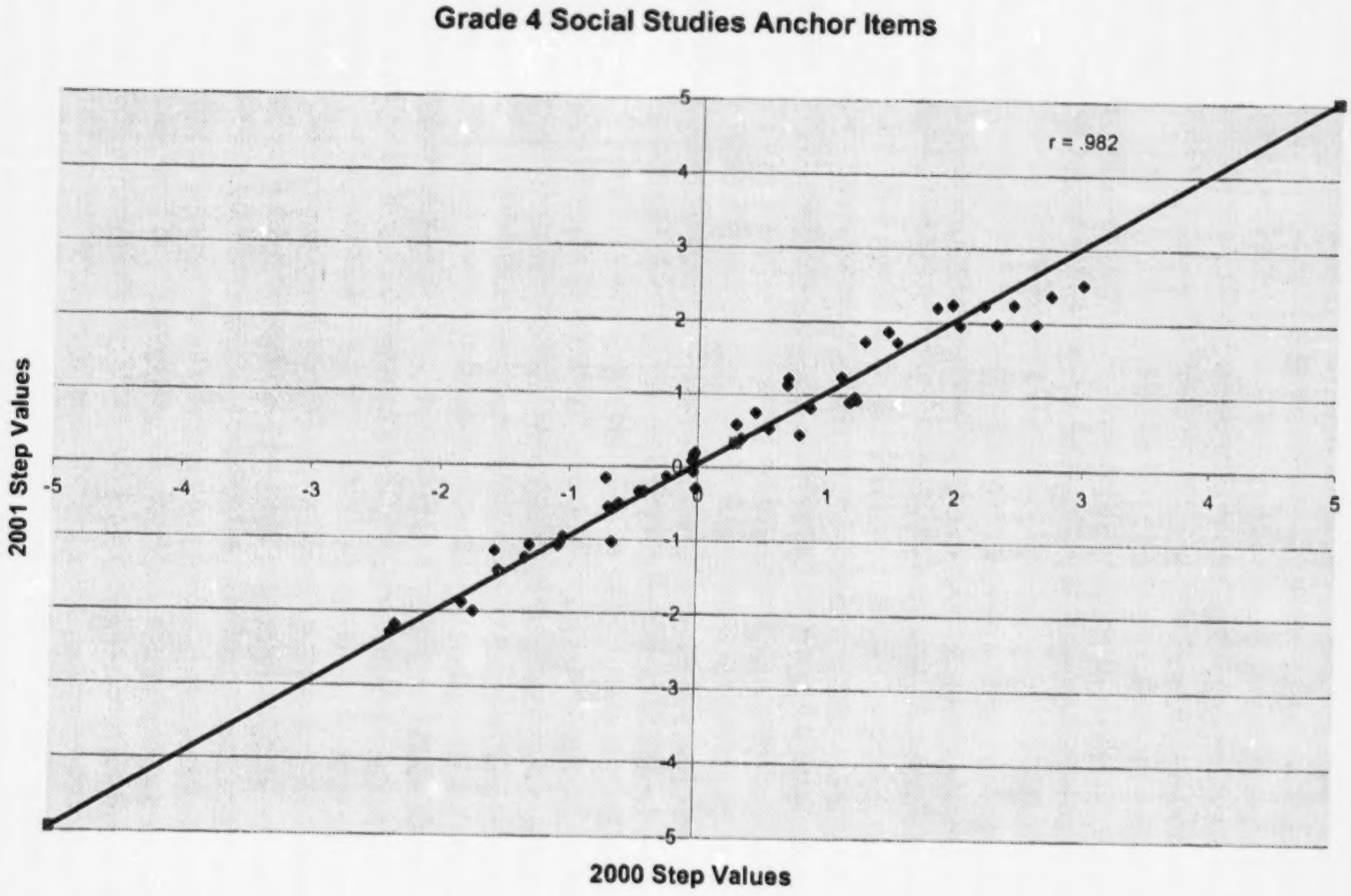

Grade 6 Science Anchor Items
r = .971
2001 Step Values
2000 Step Values
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

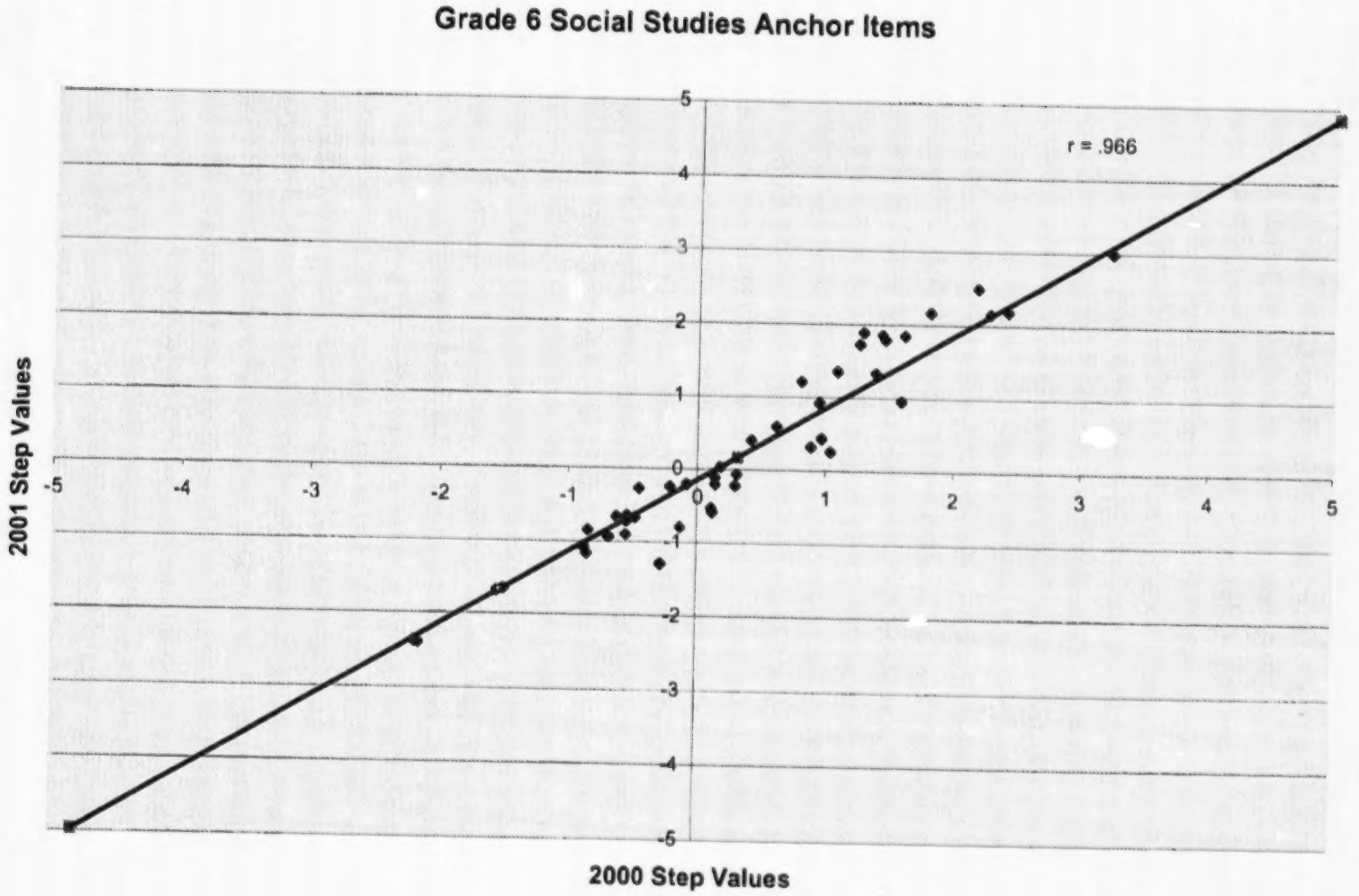


BEST COPY AVAILABLE



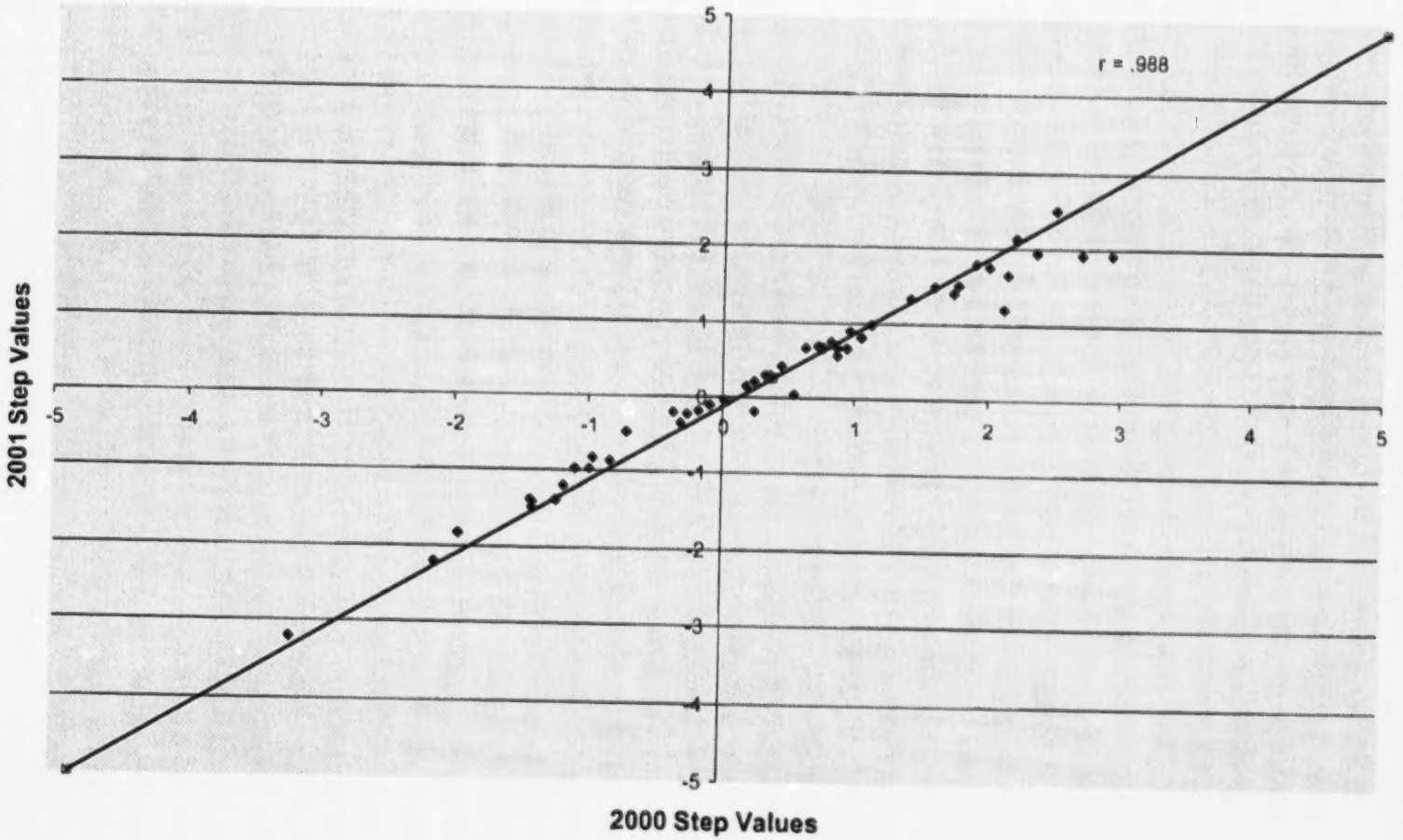
BEST COPY AVAILABLE


Grade 4 Social Studies Anchor Items
r = .982
2001 Step Values
2000 Step Values
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

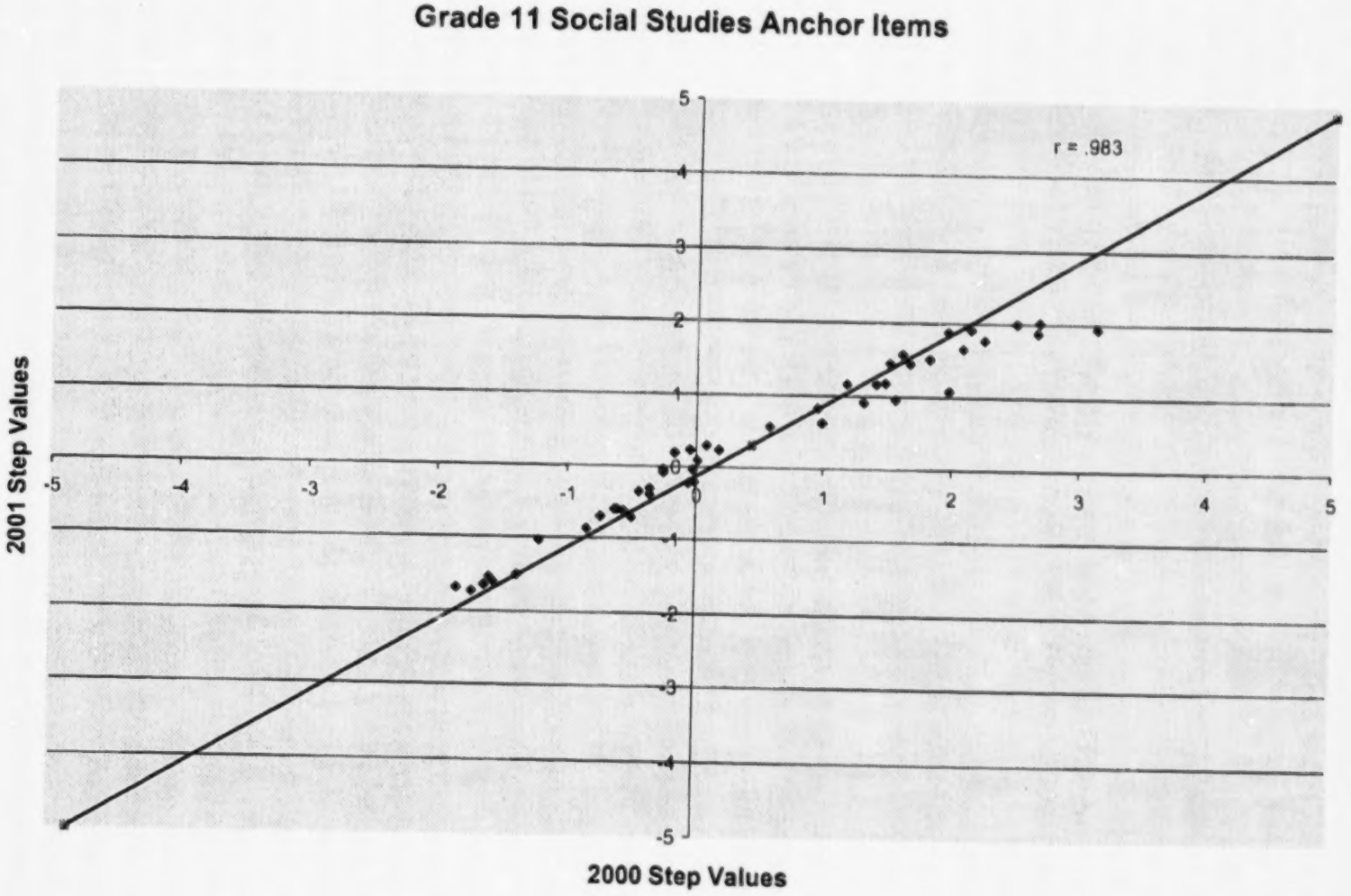


BEST COPY AVAILABLE

Grade 8 Social Studies Anchor Items



BEST COPY AVAILABLE



Attachment K

Histogram Distributions of Item Difficulties by Test and Year

Histogram Distributions of Item Difficulties by Test and Year

Note:

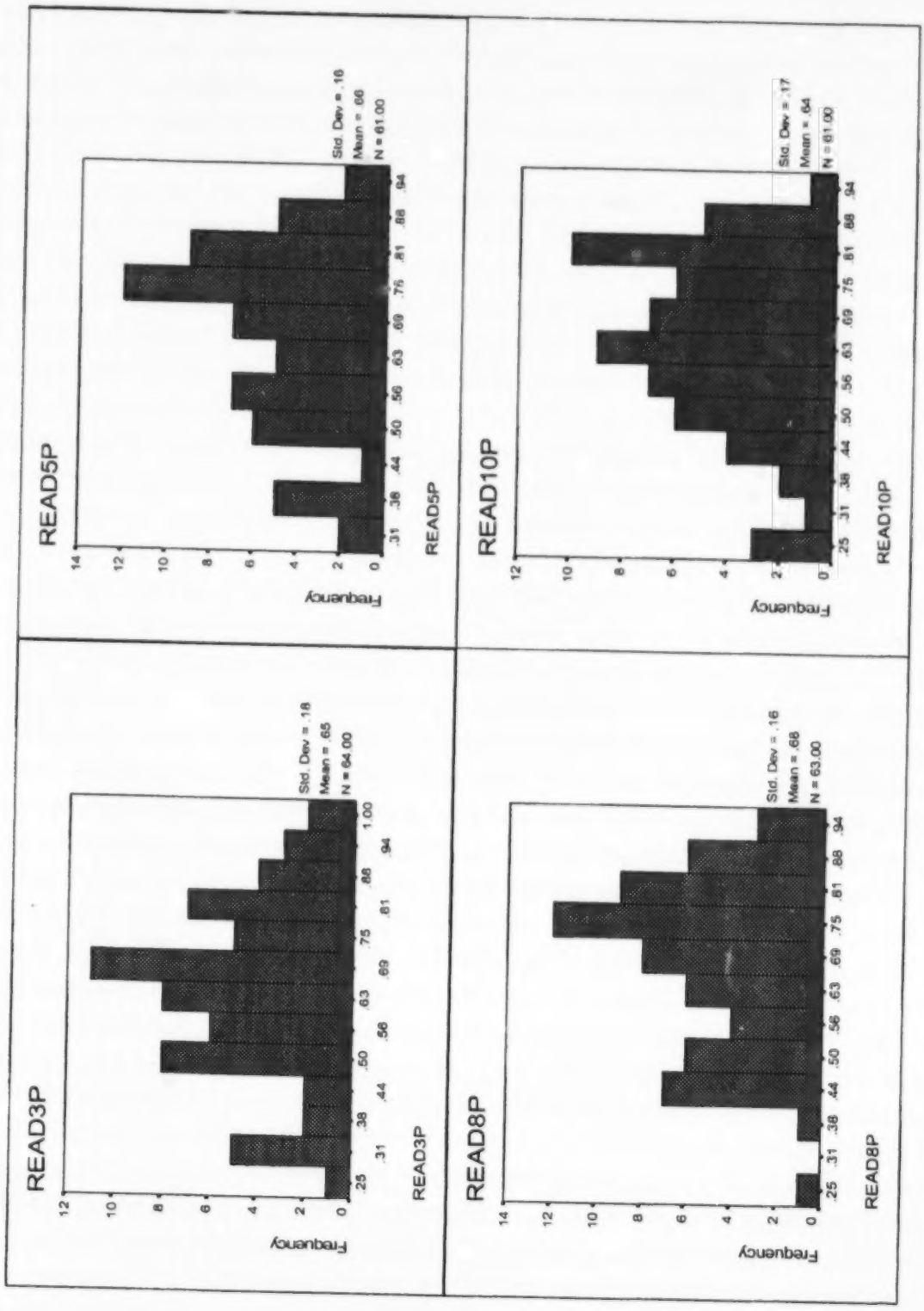
READ3P – item difficulties for Grade 3 Reading
READ5P – item difficulties for Grade 5 Reading
READ8P – item difficulties for Grade 8 Reading
READ10P – item difficulties for Grade 10 Reading
READ3B – point-biserial correlations for Grade 3 Reading
READ5B – point-biserial correlations for Grade 5 Reading
READ8B – point-biserial correlations for Grade 8 Reading
READ10B – point-biserial correlations for Grade 10 Reading

MATH3P – item difficulties for Grade 3 Math
MATH5P – item difficulties for Grade 5 Math
MATH8P – item difficulties for Grade 8 Math
MATH10P – item difficulties for Grade 10 Math
MATH3B – point-biserial correlations for Grade 3 Math
MATH5B – point-biserial correlations for Grade 5 Math
MATH8B – point-biserial correlations for Grade 8 Math
MATH10B – point-biserial correlations for Grade 10 Math

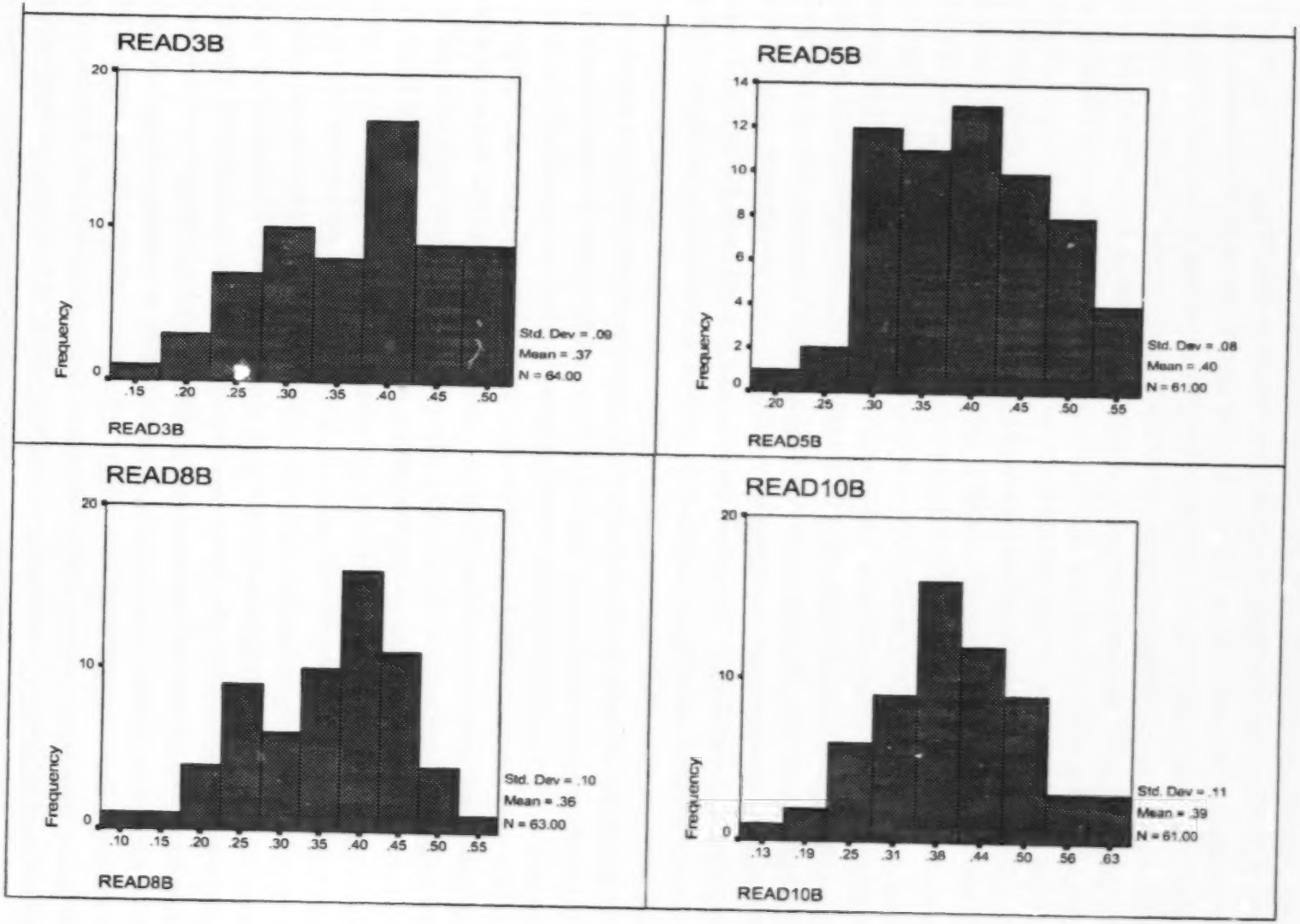
SCI4P – – item difficulties for Grade 4 Science
SCI6P — item difficulties for Grade 6 Science
SCI8P – – item difficulties for Grade 8 Science
SCI11P – – item difficulties for Grade 11 Science
SCI4B – point-biserial correlations for Grade 4 Science
SCI6B – point-biserial correlations for Grade 6 Science
SCI8B – point-biserial correlations for Grade 8 Science
SCI11B – point-biserial correlations for Grade 11 Science

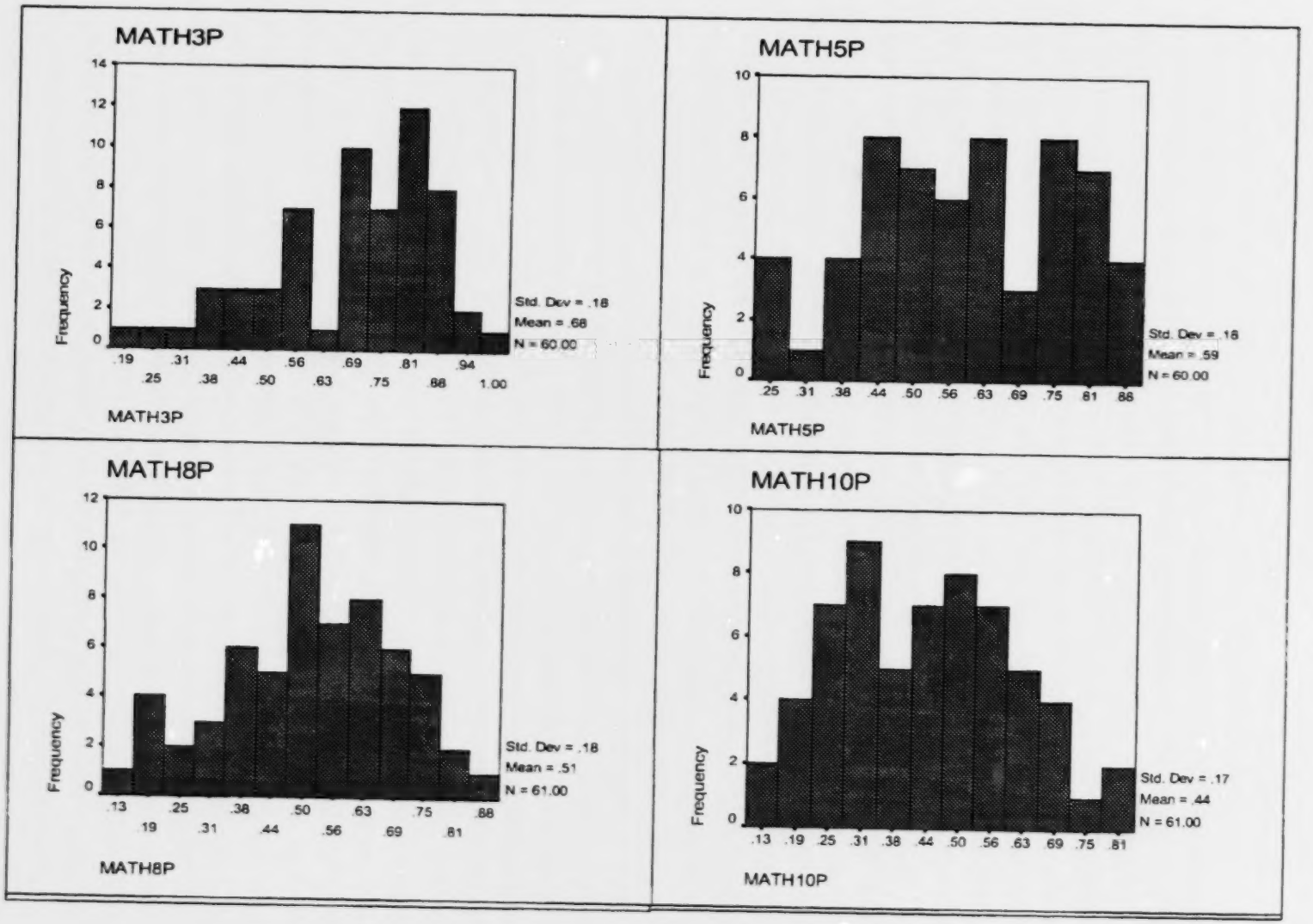
SOC4P-- item difficulties for Grade 4 Social Studies
SOC6P – item difficulties for Grade 6 Social Studies
SOC8P-- item difficulties for Grade 8 Social Studies
SOC11P – item difficulties for Grade 11 Social Studies
SOC4B – point-biserial correlations for Grade 4 Social Studies
SOC6B – point-biserial correlations for Grade 6 Social Studies
SOC8B – point-biserial correlations for Grade 8 Social Studies
SOC11B – point-biserial correlations for Grade 11 Social Studies

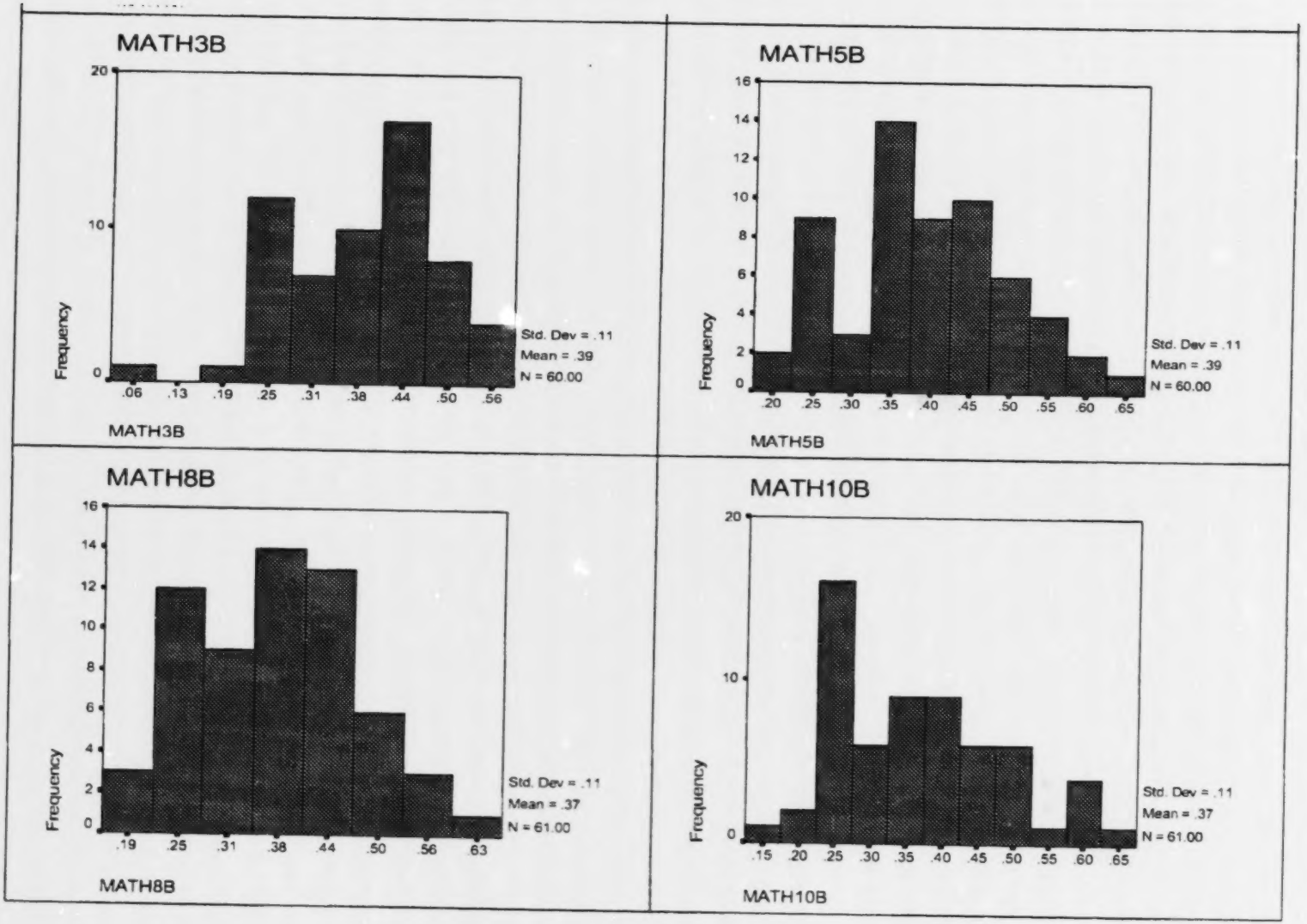
BEST COPY AVAILABLE

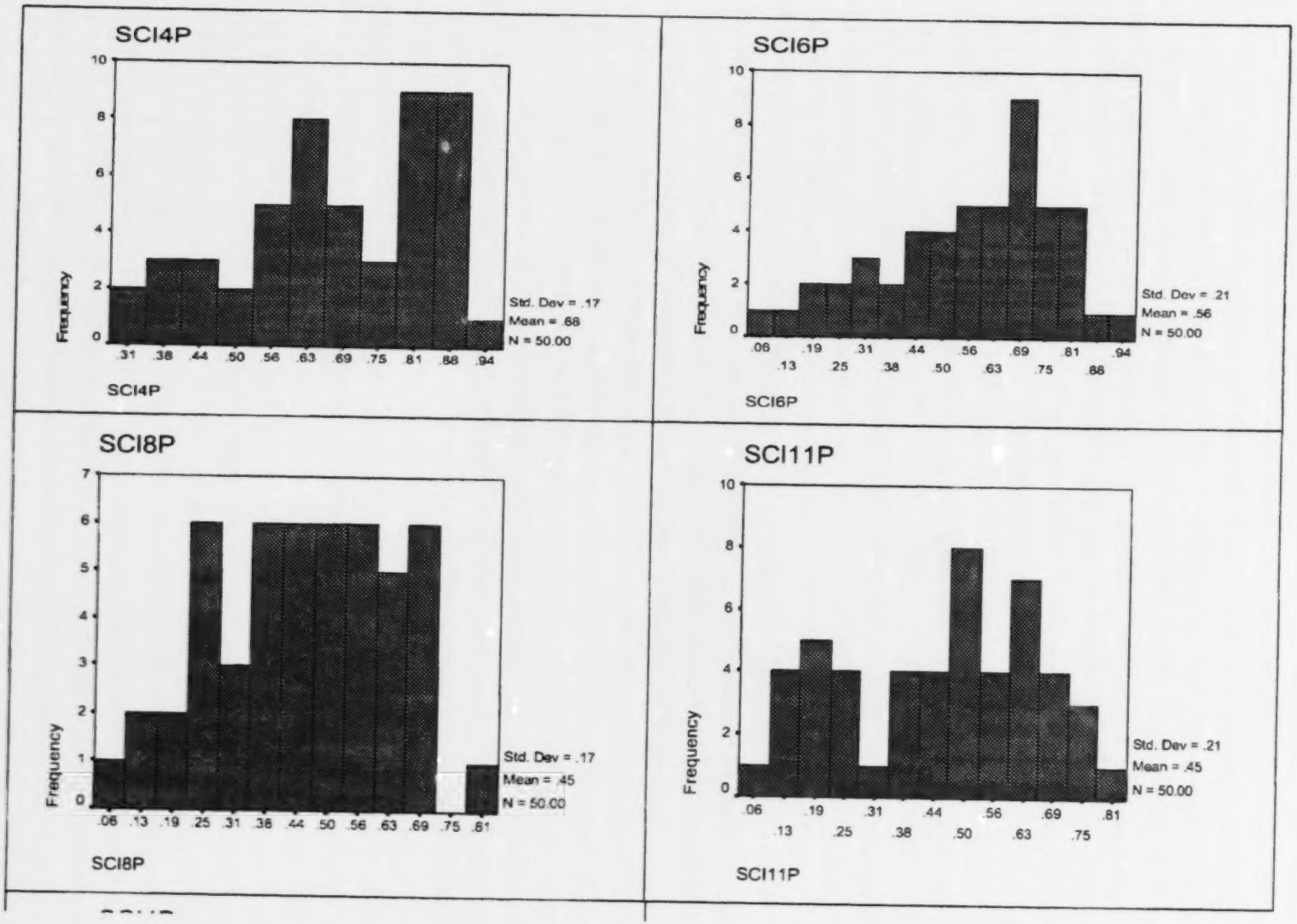


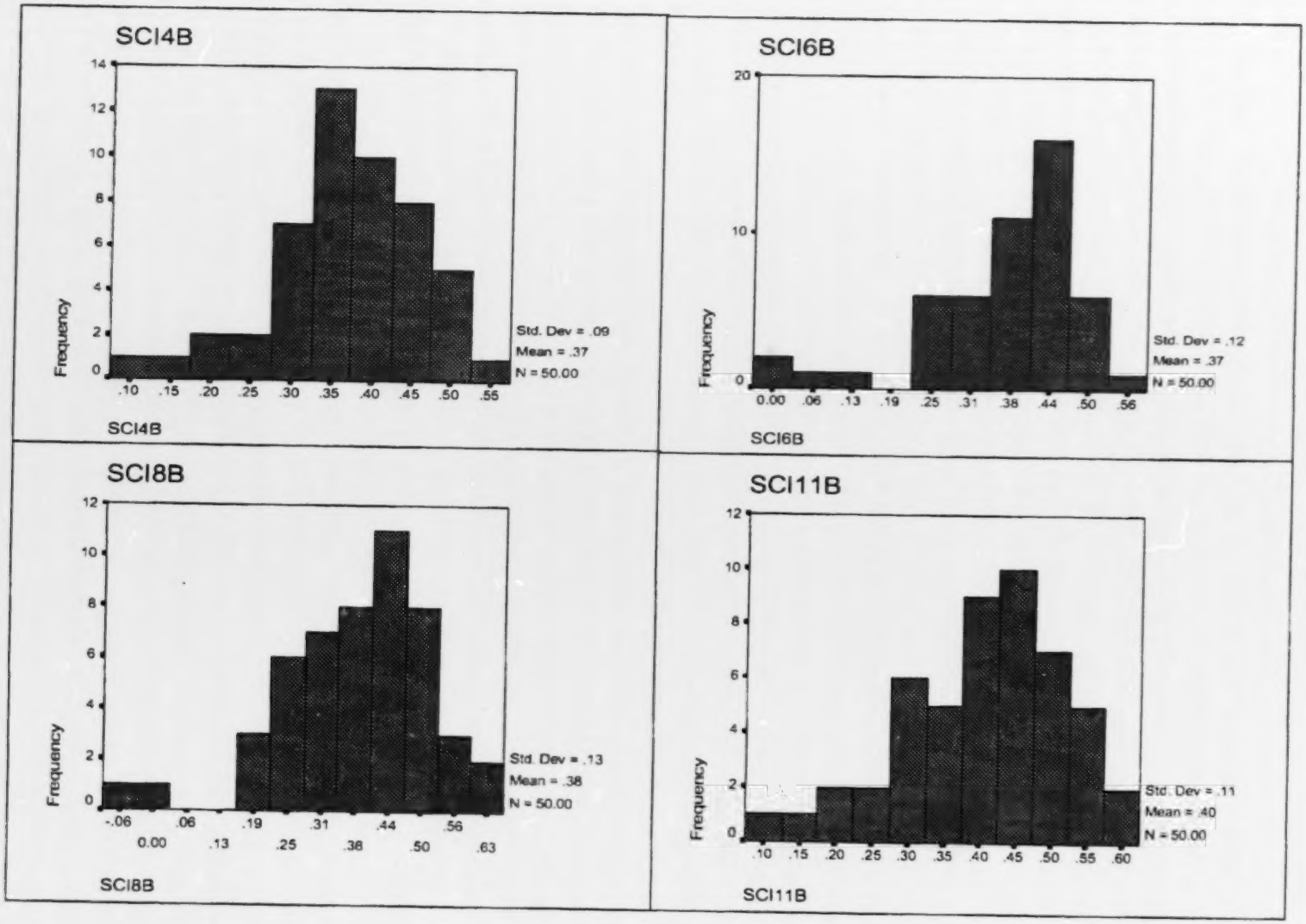
BEST COPY AVAILABLE

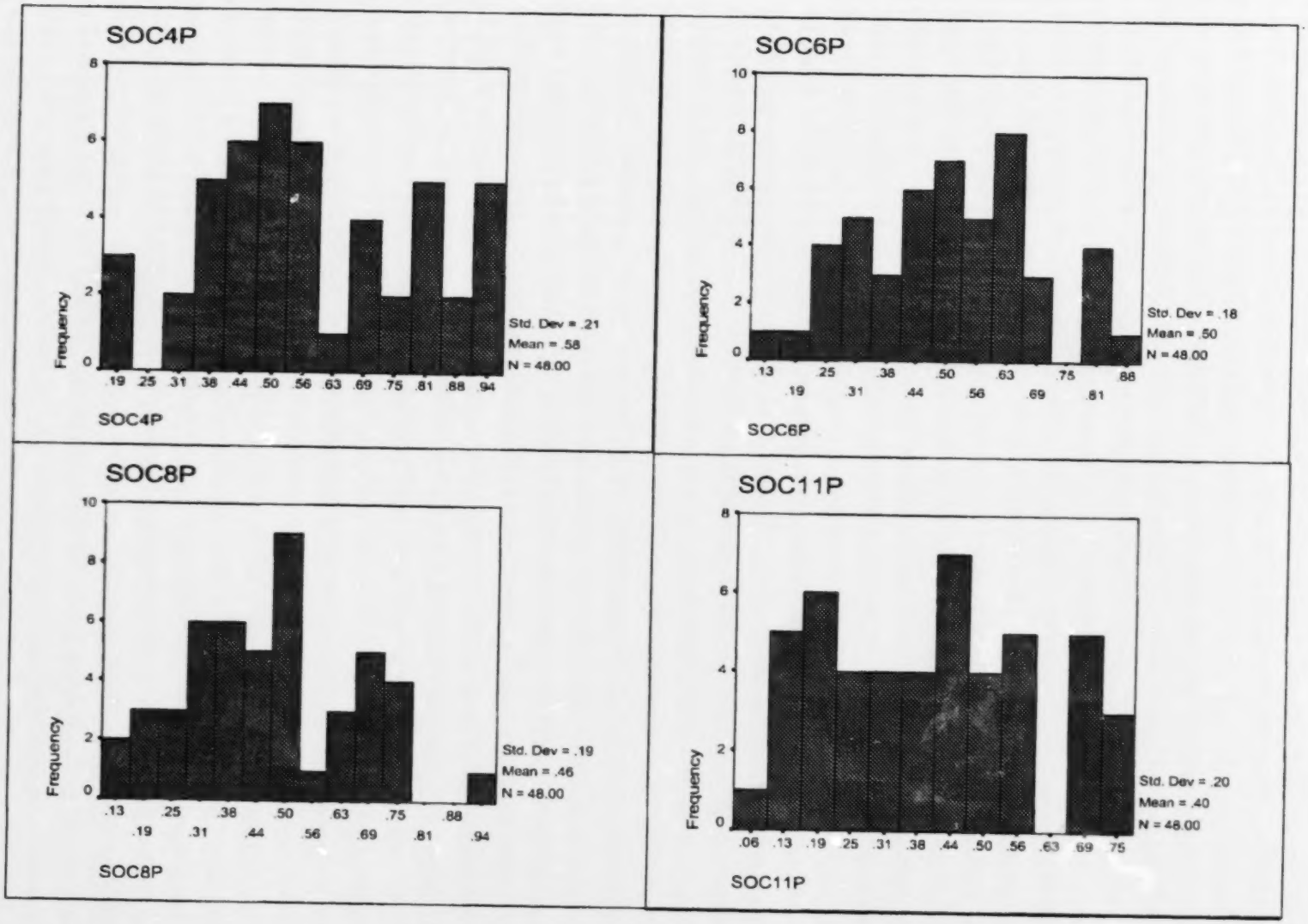

READ3B
Frequency
Std. Dev = .09
Mean = .37
N = 64.00
READ3B
READ5B
Frequency
Std. Dev = .08
Mean = .40
N = 61.00
READ5B
READ8B
Frequency
Std. Dev = .10
Mean = .36
N = 63.00
READ8B
READ10B
Frequency
Std. Dev = .11
Mean = .39
N = 61.00
READ10B

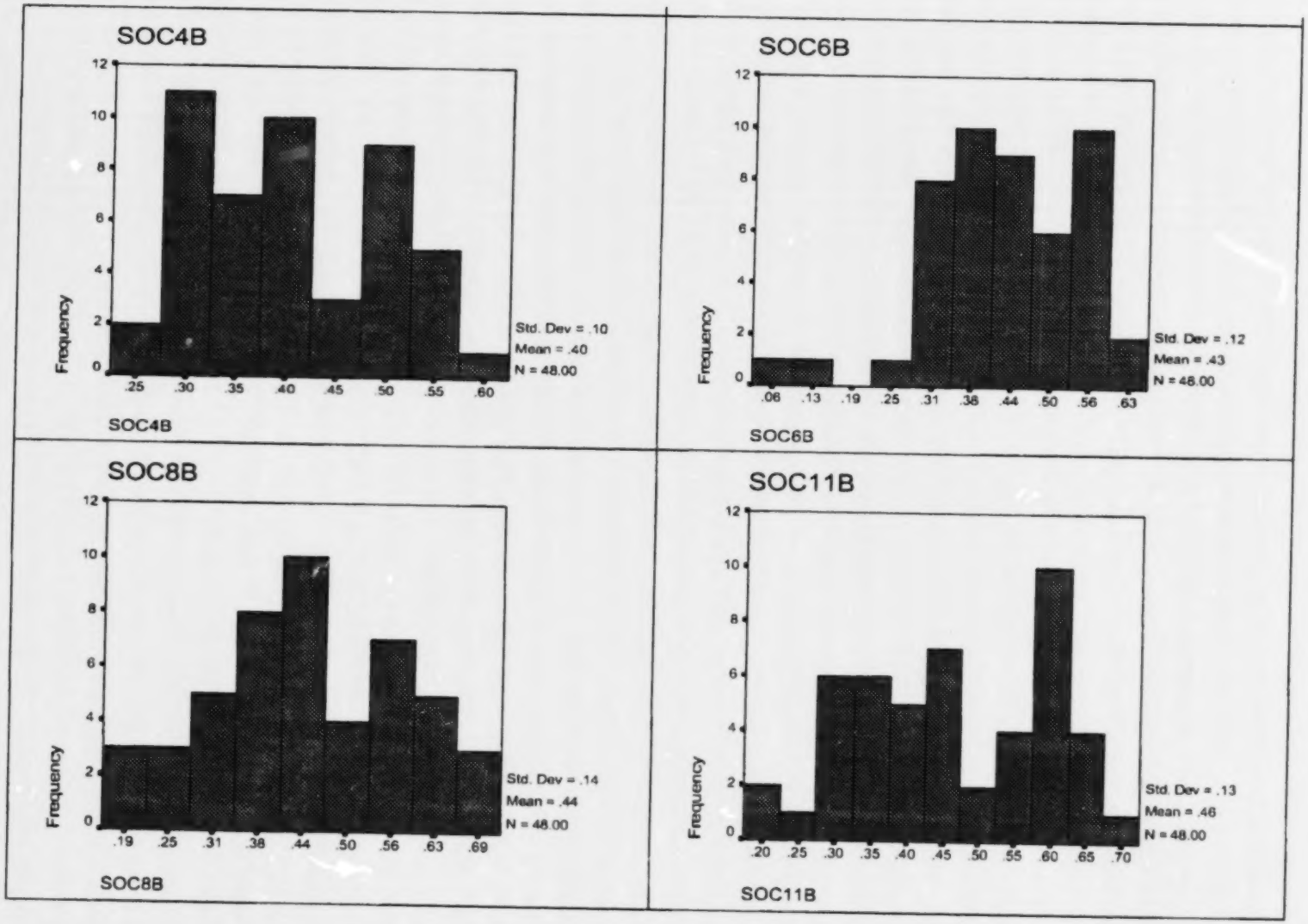

MATH3P
Frequency
Std. Dev = .18
Mean = .68
N = 60.00
MATH3P
MATH5P
Frequency
Std. Dev = .18
Mean = .59
N = 60.00
MATH5P
MATH8P
Frequency
Std. Dev = .18
Mean = .51
N = 61.00
MATH8P
MATH10P
Frequency
Std. Dev = .17
Mean = .44
N = 61.00
MATH10P


MATH3B
Frequency
Std. Dev = .11
Mean = .39
N = 60.00
MATH3B
MATH5B
Frequency
Std. Dev = .11
Mean = .39
N = 60.00
MATH5B
MATH8B
Frequency
Std. Dev = .11
Mean = .37
N = 61.00
MATH8B
MATH10B
Frequency
Std. Dev = .11
Mean = .37
N = 61.00
MATH10B


SCI4P
Frequency
Std. Dev = .17
Mean = .68
N = 50.00
SCI4P
SCI6P
Frequency
Std. Dev = .21
Mean = .56
N = 50.00
SCI6P
SCI8P
Frequency
Std. Dev = .17
Mean = .45
N = 50.00
SCI8P
SCI11P
Frequency
Std. Dev = .21
Mean = .45
N = 50.00
SCI11P


SCI4B
Frequency
Std. Dev = .09
Mean = .37
N = 50.00
SCI4B
SCI6B
Frequency
Std. Dev = .12
Mean = .37
N = 50.00
SCI6B
SCI8B
Frequency
Std. Dev = .13
Mean = .38
N = 50.00
SCI8B
SCI11B
Frequency
Std. Dev = .11
Mean = .40
N = 50.00
SCI11B


SOC4P
Frequency
Std. Dev = .21
Mean = .58
N = 48.00
SOC4P
SOC6P
Frequency
Std. Dev = .18
Mean = .50
N = 48.00
SOC6P
SOC8P
Frequency
Std. Dev = .19
Mean = .46
N = 48.00
SOC8P
SOC11P
Frequency
Std. Dev = .20
Mean = .40
N = 48.00
SOC11P


SOC4B
Frequency
Std. Dev = .10
Mean = .40
N = 48.00
SOC4B
SOC6B
Frequency
Std. Dev = .12
Mean = .43
N = 48.00
SOC6B
SOC8B
Frequency
Std. Dev = .14
Mean = .44
N = 48.00
SOC8B
SOC11B
Frequency
Std. Dev = .13
Mean = .46
N = 48.00
SOC11B

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").